AKADEMIA ŚWIĘTOKRZYSKA
IM. JANA KOCHANOWSKIEGO W KIELCACH

WYDZIAŁ MATEMATYCZNO-PRZYRODNICZY
Kierunek: Biologia

ŁUKASZ KOZŁOWSKI
Numer albumu 67668

Praca magisterska

ANALIZA FILOGENETYCZNA HISTONÓW ŁACZNIKOWYCH KRĘGOWCÓW

Promotor pracy:
prof. dr hab. Jan Pałyga

Praca przyjęta pod względem
merytorycznym i formalnym
w formie papierowej i elektronicznej

.....

KIELCE 2006

Składam serdeczne podziękowania

Panu prof. dr hab. Janowi Pałydze

za cenne wskazówki i uwagi w trakcie

pisania niniejszej pracy.

SPIS TREŚCI

WYKAZ KRĘGOWCÓW OBJĘTYCH ANALIZĄ FILOGENETYCZNĄ

Anas platyrhynchos – kaczka krzyżówka

Anguilla japonica – węgorz japoński

Anser anser anser – podgatunek gęsi gęgawy

Bos taurus – udomowione bydło

Bufo bufo gagarizans – podgatunek ropuchy szarej

Cairina moschata – kaczka piżmowa (piżmówka amerykańska)

Canis familis - pies

Carassius auratus – karaś srebrzysty (złota rybka)

Danio rerio – danio pręgowany

Gallus gallus – kura domowa

Homo sapiens – człowiek rozumny

Ictalurus punctatus – sum (sumik) kanałowy

Lytechinus pictus – gatunek jeżowca

Macaca mulatta – rezus

Mus musculus – mysz domowa

Oncorhynchus mykiss – pstrąg tęczowy

Oryctolagus cuniculus – królik

Pan troglodytes – szympans

Parechinus angulosus – gatunek jeżowca

Pongo pygmaeus – orangutan

Rattus norvegicus – szczur wędrowny

Salmo trutta fario – pstrąg potokowy

Strongylocentrotus purpuratus – jeżowiec purpurowy

Sus scrofa domestica – świnia domowa

Tetraodon nigroviridis – kolcobrzuch zielony

Xenopus laevis – żaba szponiasta

Xenopus tropicalis – gatunek żaby (brak polskiej nazwy)

SKRÓTY I SYMBOLE STOSOWANE W TEKŚCIE

B4 (H1M)	histon *Xenopus laevis* charakterystyczny dla okresu bruzdkowania
BIONJ	ulepszona wersja metody NJ
BLOSUM	macierz substytucji sekwencji białkowych
d	oczekiwana liczba substytucji nukleotydowych na miejsce dla dwóch sekwencji
F81	model substytucji nukleotydów zaproponowany przez Felsensteina w 1981 roku
FM	(Fitch-Margoliash) jedna z metod odległościowych
GTR	(general time reversible) model substytucji nukleotydów zakładający odwracalność ewolucji
H1	klasa histonów silnie lizynowych
H1°	histon H1 charakterystyczny dla komórek zróżnicowanych
H1a-e	histony H1 charakterystyczne dla komórek somatycznych
H1oo	(oocyte-specific linker histone) histon łącznikowy specyficzny dla oocytów
H1t	(testis-specific histone H1) histon H1 specyficzny dla jąder
H1X	histon *Bufo japonicus* homologiczny do B4
H2A	klasa histonów umiarkowanie lizynowych budująca rdzeń nukleosomu
H2B	klasa histonów umiarkowanie lizynowych budująca rdzeń nukleosomu
H3	klasa histonów arginowych budująca rdzeń nukleosomu
H4	histon arginowy wchodzący w skład rdzenia nukleosomu
H5	histon silnie lizynowy charakterystyczny dla jądrzastych erytrocytów
HKY	(Hasegawa-Kishino-Yano) model substytucji nukleotydów
HTH	(helix-turn-helix) motyw helisa-skręt-helisa
J69	najprostszy model substytucji zaproponowany przez Jukesa i Cantora w 1969 roku
JTT	macierz substytucji aminokwasów autorstwa Jones, Taylor, Thornton
K2P	(Kimura 2 parametr) model substytucji o dwóch parametrach

LS	(last squares) metoda ostatnich kwadratów zaliczna do metod odległościowych
ME	(minimum evolution) metoda minimalnej odległości (ewolucji)
ML	(maximum likelihood) metoda największej wiarygodności
MP	(maximum parsimony) metoda największej oszczędności
MSA	(multiple sequence alignment) dopasowanie wielu sekwencji
NJ	(neigbor joining) metoda przyłącznia najbliższego sąsiada, zaliczna do metod odległościowych
p	różnica między proporcją poszczególnych aminokwasów/nukleotydów w obrębie dwóch sekwencji
PAM	model akceptowalnych mutacji punktowych; także macierz substytucji aminokwasów oparty o ten model
PC	(Poisson corection) poprawka wartości d uwzględniająca rozkład Poissona
PDB	(Protein Data Base) bank sekwencji i struktur białkowych
p_N	substytucje niesynomiczne (zmieniające odczytu aminokwasu)
p_S	substytucje synomiczne (nie zmieniające odczytu aminokwasu)
pz	par zasad
R	współczynnik określający stosunek tranzycji do tranwersji
REV	odwracalne modele ewolucji sekwencji np. JC69, HKY, F81
Scoredist	logarytmicznie skorygowana wartość odległości d
siRNA	(small interfering RNA) mały interferujący RNA, klasa RNA o długości 21-28 par zasad
T92	model substytucji nukleotydów zaproponowany przez Tamurę w 1992 roku
UPGMA	(unweighted pair-group method using aritmetic averages) metoda nieważonych średnich połączeń zaliczna do metod odległościowych
VT	(variable time model) model substytucji aminokwasów ustalona przez Mullera i Vingron
WAG	(Whelan-and-Goldman) model substytucji aminokwasów

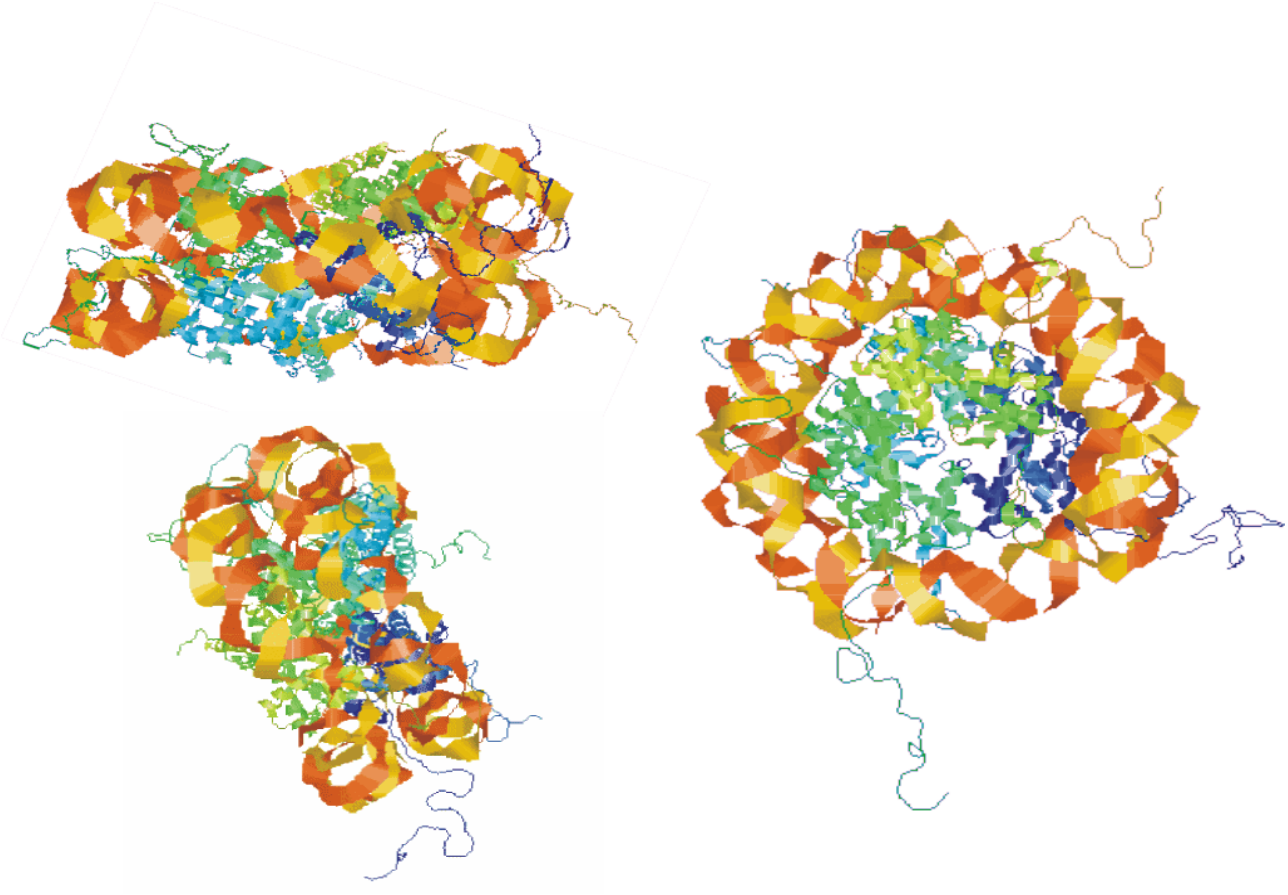
1. WSTĘP

Białka histonowe należą do wysoce konserwatywnych białek obecnych w większości organizmów eukariotycznych, a ich homologi występują również u *Procaryota* (Kasinsky i wsp., 2001). Białka te znajdują się w jądrze komórkowym, gdzie wiążą się z DNA tworząc podstawową jednostkę chromatyny jaką jest nukleosom. Struktura ta składa się z ośmiu cząsteczek histonów stanowiących rdzeń na który niemal dwukrotnie obwinięty jest DNA o długości ≈ 146 pz. Histony można podzielić na dwie podrodziny białek. Są to histony rdzeniowe i histony łącznikowe. Do histonów rdzeniowych zaliczane są białka H2A, H2B, H3 i H4. Każde z nich występuje podwójnie w nukleosomie tworząc tetramer $H3_2$-$H4_2$ i dwa dimery H2A-H2B, które z kolei łączą się w oktamer (Rys. 1). Białka te mają budowę modularną i składają się z trzech części: COOH-terminalnej, centralnej i NH_2-terminalnej. Rdzeń nukleosomu budują jedynie dwie pierwsze domeny, podczas gdy części N-terminalne wystają na zewnątrz nukleosomu. Pomimo swej konserwatywności ewolucyjnej (histony należą do najwolniej się zmieniających białek) histony wykazują pewne zróżnicowanie, jedynie histon H4 występuje w formie jednorodnej. Drugą podrodzinę stanowią histony łącznikowe obecne w większej liczbie wariantów i szybciej ewoluujące niż pozostałe histony. Białka te ulokowane są w miejscu w którym DNA wchodzi do i wychodzi z nukleosomu stanowiąc swego rodzaju klamrę spinającą całość (Ramakrishnan, 1997; Kłyszejko-Stefanowicz, 2002; Luger i Hansen, 2005; Chakravarthy i wsp., 2005)

1.1. BUDOWA HISTONÓW ŁĄCZNIKOWYCH

Histony łącznikowe to małe, silnie zasadowe białka o masie około 21 tys. Da. Zaliczamy tu histony H1 powszechnie występujące we wszystkich rodzajach komórek jądrzastych oraz histony H5 specyficzne dla jądrzastych erytrocytów ptaków i płazów. Białka te podobnie jak histony rdzeniowe zbudowane są z trzech domen. Część centralna (zwana też globularną) zbudowana z około 80 aminokwasów

cechuje się najwyższą konserwatywnością (Wierzbicki, 2002; Kasinsky i wsp., 2001). Domena globularna składa się z wiązki trzech α-helis oraz β-harmonijki (nazywanej w tym przypadku skrzydłem) ulokowanej w pobliżu C-końca. Ze względu na taką budowę histony łącznikowe zaliczyć można do rodziny białek HTH, choć typowe białka HTH takie jak CAP (białko aktywatora katabolicznego) mają między drugą a trzecią helisą skręt zbudowany z czterech aminokwasów, którego brak u histonów (Ramakrishnan, 1997). Motyw uskrzydlonej helisy decyduje o możliwościach wiązania się do DNA, a różnice w składzie aminokwasowym między histonami H1 i H5 skutkują odmiennym powienowactwem tych form do DNA (Gajiwala i Burley, 2000).



Rys. 1 Nukleosom (kod PDB: 1KX5). Model oparty na podstawie krystalografii rengenowskiej o rozdzielczości 1.9 Å zobrazowano za pomocą programu RasMol (Bernstein, 2000). DNA zaznaczono kolorem żółtym i czerwonym (na obrzeżach), białka histonowe zaznaczono na niebiesko, zielono (w centrum). Z nukleosomu wystają N-terminalne części histonów rdzeniowych. Nukleosom u góry po lewej pokazano od strony w której DNA wchodzi i schodzi.

Część N-terminalna zbudowana jest z 35-40 aminokwasów z dużą ilością aminokwasów zasadowych oraz proliny i alaniny. Dodatkowo domenę tą można podzielić na dwie części: zewnętrzną silnie hydrofobową i sąsiadującą z domeną centralną część zasadową. W obrębie struktury drugorzędowej można tu wyróżnić dwie α-helisy oddzielone dwoma glicynami dzięki którym motyw helisa-Gly-Gly-helisa jest elastyczny. Ponadto helisy te są silnie alifatyczne czyli aminokwasy zasadowe położone są z jednej strony, zaś zasadowe z drugiej oraz zawierają potrójne miejsca zasadowe co jest cechą białek wiążących DNA takich jak protaminy (Vila i wsp., 2002). Należy podkreślić, że wyżej opisana struktura drugorzędowa dotyczy jedynie części zasadowej i występuje jedynie w obecności DNA (Vila i wsp. 2001). Domena COOH-końcowa histonu H1 jest najdłuższa i zbudowana jest z 90-160 aminokwasów wśród których przeważają lizyna, arginina i prolina, które stanowią ponad 85% aminokwasów histonu H1 co powoduje niemal 15-krotną przewagę aminokwasów zasadowych nad kwasowymi (Kłyszejko-Stefanowicz, 2002). Dodatkowo często występują seryna i treonina będące miejscem fosforylacji. Innym specyficznym motywem są sekwencje (S/T)PXX, gdzie X oznacza lizynę lub argininę. Motyw ten decyduje o możliwości wiązania się do mniejszego rowka DNA (Ramakrishnan, 1997). Podobnie jak domena N-końcowa ogon C-terminalny nie wykazuje struktury drugorzędowej, chyba, że połaczy się on z DNA lub będzie stabilizowany przez specjalne związki takie jak trifluoroetanol czy $NaClO_4$. W takich warunkach powstaje kilka alifatycznych α-helis przedzielonych β-skrętem lub σ-skrętem dzięki czemu histony łącznikowe mogą się wiązać zarówno z mniejszym jak i większym rowkiem DNA (Vila i wsp., 2000).

Tabela 1. Średnia długość poszczególnych domen histonu H1 u kręgowców.

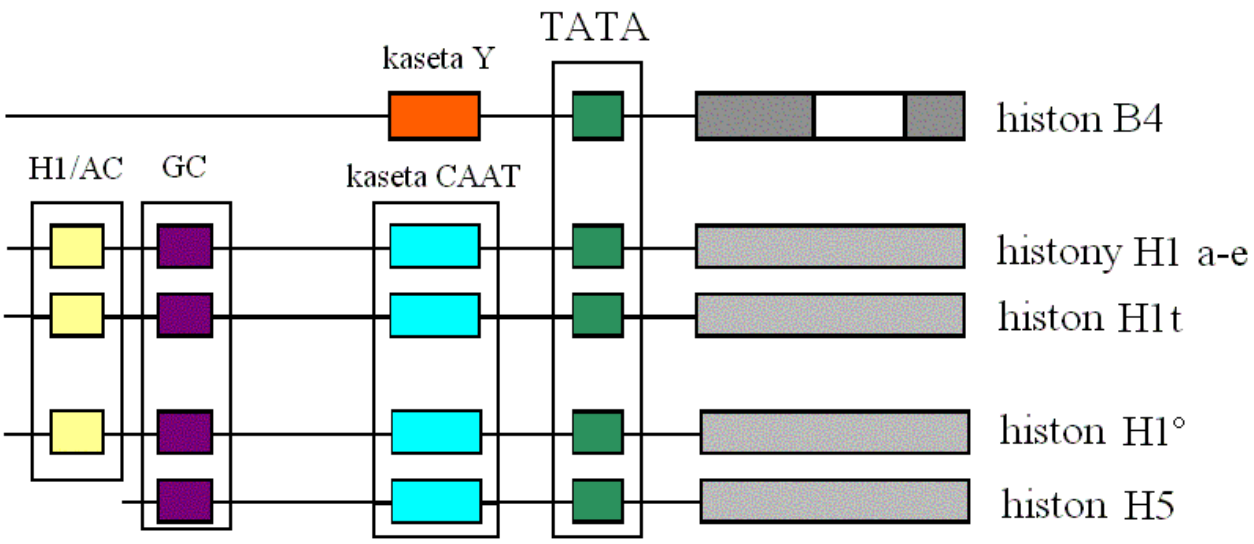
Ogon NH_2-terminalny	Domena globularna	Ogon COOH-terminalny
40±13	79±5	106±17

1.2. FUNKCJA HISTONÓW ŁĄCZNIKOWYCH

Podstawową funkcją histonu H1 i H5 jest stabilizowanie zwartej struktury nukleosomu. Brak histonów łacznikowych zmienia budowę chromatyny, która staje się luźna, a nukleosomy ją budujące pozbawione są charakterystycznej struktury łodyżki. Uniemożliwia to wykształcenie następnego stopia konformacji przestrzennej jakim jest 30 nm włókno, które przez wiele lat opisywano jako solenoid, ale obecnie coraz częściej przedstawia się według modelu wstęgi zygzakowatej (Bednar i wsp., 1998; Travers, 1999). Oprócz tej wydawałoby się czysto mechanicznej funkcji histony łacznikowe podlegają licznym modyfikacją takim jak fosforylacja, ubikwitynacja, acetylacja, ATP-rybozylacja i metylacja co ma duże znaczenie na regulację ekspresji genów (Kłyszejko-Stefanowicz, 2002). Przykładowo fosforylacja motywu SPKK w C-terminalnej części wpływa na zdolność wiązania się histonu H1 do chromatyny, a siła tego oddziaływania zależy także od lokalizacji danego motywu co sugeruje, że poszczególne sekwencje ulegające modyfikacji nie są równocenne (Hendzel i wsp., 2004). Obecnie zebrane dane jasno wskazują, że funkcja histonów łącznikowych jest daleka od początkowo postulowanej jaką była rola generalnego represora espresji genów. Można tu zaobserwować istnienie różnych sprzężeń zwrotnych, przykładowo wyciszenie za pomocą siRNA histonów H1 u *Arabidopsis thaliana* powoduje odziedziczalne zmiany we wzorze metylacji DNA i ekspresji genów (Wierzbicki i Jerzmanowski, 2004) Sprawa jest na tyle skomplikowana, że postuluje się nawet istnienie specjalnego kodu według którego poszczególne modyfikacje w odmienny sposób wpływają na ekspresję genów, kondensację chromatyny i inne procesy komórkowe (Jenuwein i Allis, 2001; Turner, 2002).

1.3. ZRÓŻNICOWANIE HISTONÓW ŁĄCZNIKOWYCH

Histony H1 i H5 są białkami najbardziej zróżnicowanymi w całej rodzinie histonowej. Wszystkie do tej pory zbadane organizmy (ponad 100 gatunków roślin i zwierząt) posiadają więcej niż jeden wariant tego białka (Sullivan, 2002).



Rys. 2 Podział histonów łącznikowych uwzględniający strukturę genów. Szary prostokąt oznacza sekwencję kodującą (egzon), biały sekwencję nie kodującą (intron), na zielono zaznaczono kasetę TATA, na niebiesko kasetę CAAT, kolor fioletowy oznacza region bogaty w nukleotydy GC, żółty sekwencję specyficzną dla H1 (kaseta bogata w AC; 5'AAGAAACACACA3'), a na pomarańczowo oznaczono kasetę Y (sekwencja odwrócona w stosunku do CAAT; na podstawie Khochbin i Wolffe, 1994; zmienione).

Najdokładniej zbadano pod tym kątem ssaki u których można wyróżnić co najmniej osiem wariantów: pięć somatycznych (H1a-e) oraz histony H1t, H1oo i $H1^0$ (Rys. 2). Pod względem budowy genów i białek oraz czasu i miejsca ekspresji histony łącznikowe można podzielić na cztery klasy. Są to:

- warianty, których ekspresja następuje we wczesnym etapie rozwoju;
- warianty występujące w komórkach somatycznych;
- histony łącznikowe związane z procesem różnicowania;
- wariant charakterystyczny dla jąder.

Histony pierwszej grupy ulegają ekspresji jedynie w intensywnie proliferujących komórkach w okresie oogenezy i na początku embrionezy, a dokładniej mówiąc w czasie bruzdkowania (ang. cleavage linker histones). Początkowo białka te wyizolowano u płazów (wariant B4 zwany też H1M u *Xenopus laevis* oraz H1X u *Bufo japonicus*) i bezkręgowców (grupa histonów cs-H1 u *Strongylocentrotus purpuratus, Parechnus milaris* i *Lytechinus pictus*). Ich homolog u ssaków wykryto po dość żmudnych poszukiwaniach w oocytach myszy i nazwano

H1oo (Tanaka i wsp., 2001). W obrębie promotora brak u nich kasety bogatej w guaninę, kasety charakterystycznej dla H1 oraz kasety CAAT (zamiast niej jest kaseta Y). Sam odcinek kodujący posiada introny, których inne histony nie posiadają, a na końcu odcinka 3' znajduje się sygnał poliadenylacji (Khochbin i Wolffe, 1994; Khochbin, 2001). Odmienność histonów okresu bruzdkowania widoczna jest również na poziomie białka, gdyż są one znacznie dłuższe (273-285 aminokwasów) i słabiej zasadowe (proporcje aminokwasów kwaśnych i zasadowych są niemal równe) od innych histonów łącznikowych (Tanaka i wsp., 2001). Skutkuje to słabszym powinowactwem histonów B4, H1oo i cs-H1do DNA co może być cechą korzystną w komórkach podlegających częstym replikacjom.

Somatyczne histony łącznikowe nazywane są histonami zależnymi od replikacji, ponieważ ich synteza zachodzi jedynie w fazie S występują w większości komórek w tym także w komórkach embrionalnych i oocytach (Clarke i wsp., 1992; Clarke i wsp., 1997). Sekwencje promotorowe zawierają bogatą w guaninę, kasety TATA i CAAT oraz enhancer zlokalizowany 450-480 pz powyżej czapeczki (tzw. sekwencja bogata w TG; Khochbin, 2001). Somatyczne histony łącznikowe, choć ulegają ekspresji we wszystkich komórkach wykazują różne proporcje. Przykładowo w komórkach proliferujących przeważają histony H1a i H1b, zaś w komórkach zróżnicowanych warianty H1c, H1d i H1e. Podobne zróżnicowanie można zaobserwować w odniesieniu do poszczególnych tkanek (Lennox i Cohen, 1983). Różnice te zależą m.in. od poziomu ekspresji genów, okresu półtrwania mRNA i białek (Wang i wsp., 1997). Różnice sekwencji aminokwasów, które wahają się od 15-40% i powodują odmienne powinowactwo poszczególnych wariantów do DNA i chromatyny. Inne są także miejsca, które ulegają modyfikacją posttranslacyjnym.

Przez histony związane z procesem różnicowania należy rozumieć warianty $H1^0$ i H5. Pierwszy występuje w zróżnicowanych komórkach somatycznych (mózg, wątroba, płuca), drugi zaś jest charakterystyczny dla jądrzastych erytrocytów. Posiadają one wspólną budowę promotora (Rys. 2) z kasetami H1/AC, GC i nietypową kasetą TATA, brak natomiast sekwencji CAAT zamiast której występuje sekwencja spotykana w promotorze histonu H4 mianowicie 5'TCANNNNGGTCC'3

będąca miejscem wiązania się specyficznego czynnika transkrypcji H4TF2 (dla prostoty pominięto ten fakt na rysunku). Wszystkie te elementy są konserwatywne i silnie oddziaływują na ekspresję genu. Innym elementem kilkakrotnie pojawiającym się w dalszej cząsci promotora wpływającym w znaczny sposób na poziom ekspresji histonu $H1^0$ jest sekwencja (A/C)GGGGGG(A/C) nazywana ścieżką heksadeoksyguaninową $(dG)_6$ (Khochbin i Wolffe, 1994; Dong i wsp., 1995). Dominującą rolę w kontroli ekspresji genu H5 mają trzy elementy promotora: region bogaty w GC, miejsce wiążące USF (ang. USF binding side) oraz element UPE (ang. upstream positive element) zawierający w sobie wyżej wyminiony element charakterystyczny dla histonu H4. Na poziomie mRNA histony te wyróżniają się długimi odcinkami nie podlegającymi translacji flankującymi oba końce transkryptu (5'-UTR i 3'-UTR; Franke i wsp., 1998). Powstające z nich białka zawierają dużo seryny, alaniny i lizyny (wariant H5) lub argininy (wariant $H1^0$). Ponadto występuje tu niespotykana u innych histonów łącznikowych treonina i histydyna (Kłyszejko-Stefanowicz, 2002). Jak wspomniano histony ulegają licznym modyfikacją, a jedna z nich jest zarezerwowana właśnie dla histonu $H1^0$ w obrębie którego możemy wyróżnić dwie frakcje w zależności od tego czy histon $H1^0$ jest acetylowany na N-końcu czy nie. Ilość histon $H1^0$ i histonu H5 pozytywnie koreluje z kondensacją chromatyny oraz spadkiem replikacji i transktypcji co związane jest z wysokim powinowactwem tych wariantów do DNA (Koutzamani i wsp., 2002).

Ostatnią wydzieloną grupę w obrębie histonów łącznikowych tworzy histon H1t specyficzny dla męskich komórek rozrodczych. Transkrypt H1t wykrywa się w stermatocytach I rzędu znajdujących się w profazie I mejozy (środkowy i późny pachyten), zaś produkt białkowy obecny jest do początkowego okresu rozwoju spermatyd, gdzie może stanowić ponad 50% histonu H1. W późniejszym etapie rozwoju komórek plemnikowych histony zostają zastąpione kolejno przez białka TP (ang. transition proteins) i protaminy (Grimes i wsp., 2003). Mimo podobnej budowy promotora jak u wariantów H1a-e histon H1t ulega specyficznej tkankowo i czasowo ekspresji co tłumaczy się istnieniem negatywnych elementów regulatorowych (Wilkerson i wsp., 2002). mRNA nie posiada ogona poli-(A), w miejscu którego

znajduje się konserwatywna 26 nukleotydowa sekwencja tworząca strukturę „szpiki do włosów” (ang. hairpin loop), która odpowiada za dojrzewanie i stabilizację mRNA (Dominski i Marzuluff, 1999). Na poziomie białka wariant ten cechuje obecność metioniny, której brak u innych histonów H1 (Fantz i wsp., 2001)

Oprócz wymienionych wariantów nieallelicznych powstałych prawdopodobnie w wyniku duplikacji genów i ich późniejszej specjalizacji w miarę nagromadzenia się różnic w obrębie sekwencji nukleotydowej histony łącznikowe wykazują polimorfizm w obrębie alleli co opisano u wielu gatunków ssaków (Pałyga, 1990) i ptaków (Pałyga i wsp., 2000; Kowalski i wsp., 2004).

1.4. INAKTYWACJA HISTONÓW H1 A ICH ZNACZENIE

Ze względu na swoje powszechne występowanie i istotną funkcję trudno sobie wyobrazić, aby histony łącznikowe były białkami, których inaktywacja może nie mieć żadnych konsekwencji. Jednak początkowe badania polegające na tworzeniu transgenicznych myszy pozbawionych jednego lub nawet kilku wariantów histonów łącznikowych były pod tym względem dość zaskakujące. Okazało się, że brak jednego wariantu był kompensowany przez inne formy histonów łącznikowych, a zwierzęta nie wykazywały żadnych zauważalnych zmian (Sirotkin i wsp. 1995). Dotyczyło to nawet silnie wyspecjalizowanych wariantów takich jak H1t. Myszy pozbawione tego wariantu (H1t -/-) były w pełni żywotne i co dosyć zaskakujące płodne, również potomstwo nie wykazywało żadnych odchyleń od normy (Fantz i wsp., 2001; Lin i wsp., 2000). Podobnie brak dwóch wariantów nie powodował większych zmian (Fan i wsp., 2001), choć w tym przypadku da się już wykryć subtelne zmiany na poziomie ekspresji niektórych genów, które paradoksalnie w większości polegały na jej obniżeniu (Alami i wsp., 2003). Dla kontrastu brak trzech wariantów histonów łącznikowych powoduje całkowitą letalność na poziomie embrionalnym (Fan i wsp., 2003; Fan i Skoultchi, 2003). Podobne badania przeprowadzone na linii komórkowej DT40 kury przyniosły analogiczne wyniki (Takami i Nakayama 1997; Takami i wsp., 2000).

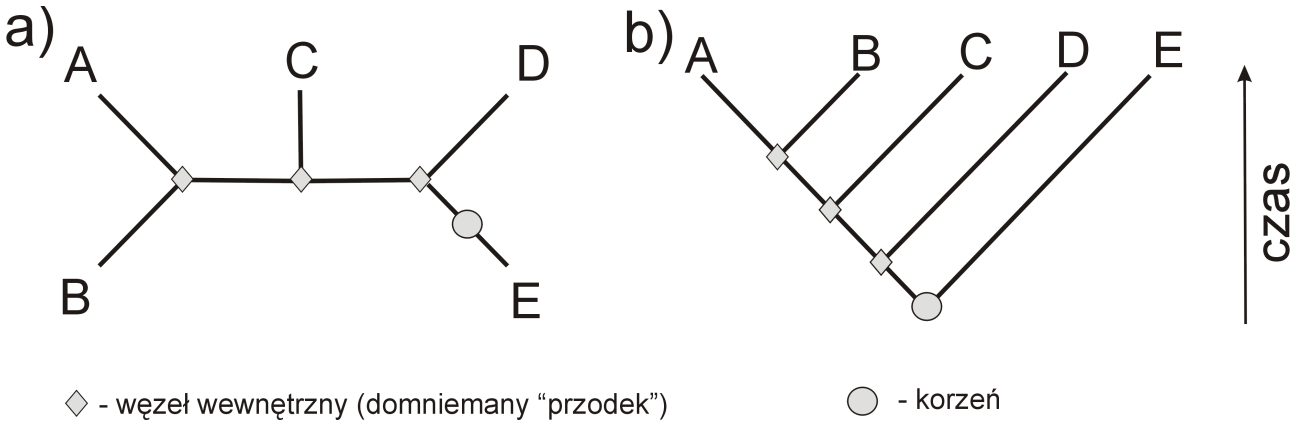
Bardziej szczegółowe informacje dotyczące struktury i funkcji histonów łącznikowych można znaleźć w innych źródłach (Kozłowki, 2004).

1.5. ANALIZA FILOGENETYCZNA HISTONÓW ŁĄCZNIKOWYCH

Celem niniejszej pracy jest przeprowadzenie analizy filogenetycznej białek H1 i H5 u kręgowców w związku z czym na początek zostaną przedstawione podstawowe wiadomości na temat filogenetyki molekularnej. Badania te mają zbadać stosunki między genami rodziny histonów łącznikowych, które ukształtowały się w czasie ich ewolucji.

2. PODSTAWY FILOGENETYKI MOLEKULARNEJ

Filogenetyka to dział biologi zajmujący się badaniem relacji między organizmami żywymi w oparciu o podobieństwa w budowie sekwencji nukleotydowych i białkowych. Opiera się ona na założeniu, że wraz z upływem czasu następują zmiany (mutacje) w obrębie materiału genetycznego (DNA), które odzwierciedlają historię organizmu i jego potomków. Pokrewieństwo między analizowanymi sekwencjami przedstawia się w postaci grafu zwanego drzewem.

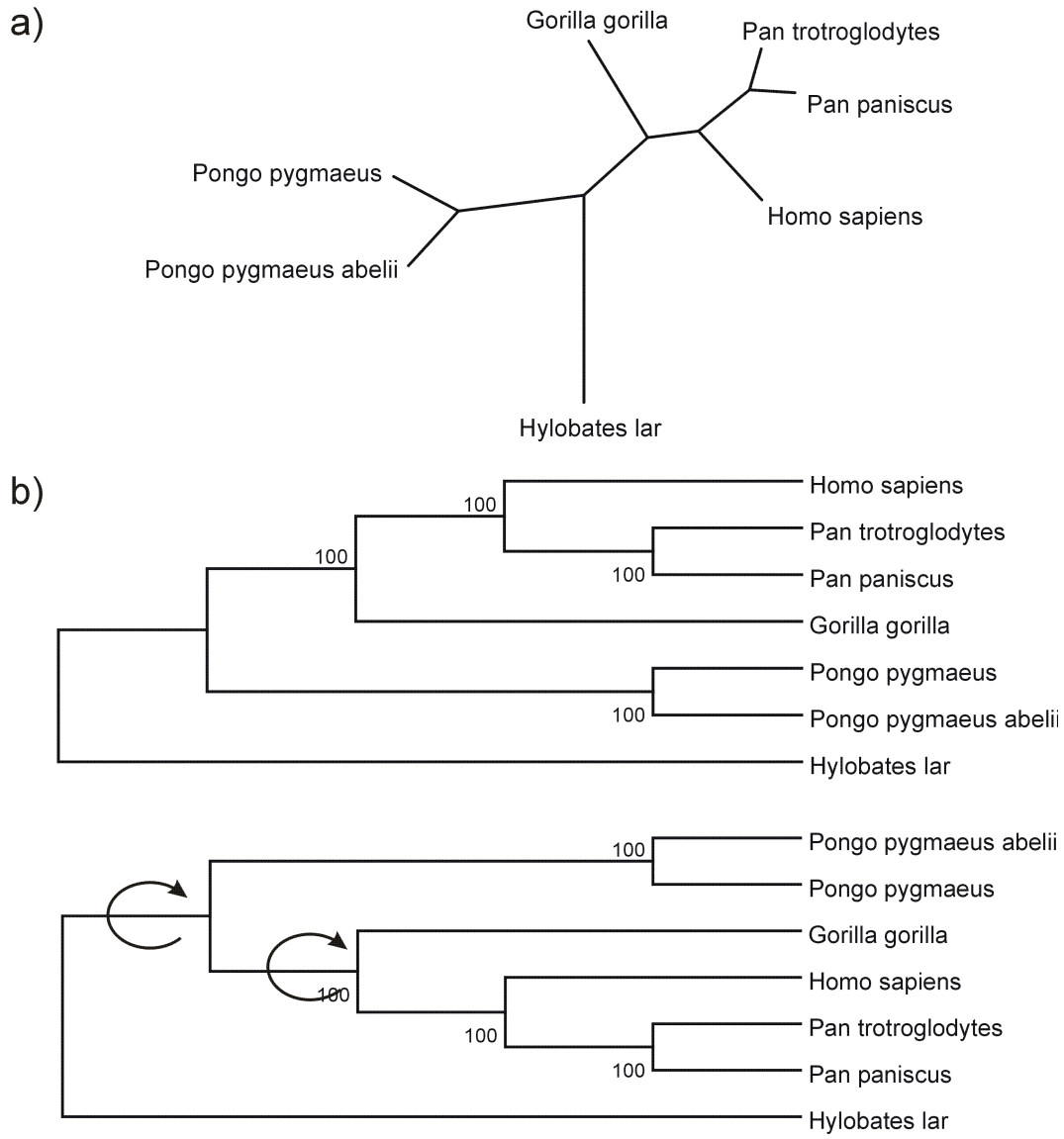


Rys. 3 Typy drzew. a) drzewo nieukorzenione; b) drzewo ukorzenione. Poszczególne litery symbolizują badane sekwencje, będące jednocześnie węzłami zewnętrznymi (liśćmi) drzewa. Odcinki między poszczególnymi węzłami nazywamy gałęziami, tak więc węzły są miejscami rozwidlenia rozgałęzień.

2.1. DRZEWA FILOGENETYCZNE

Istnieją dwa rodzaje drzew ukorzenione i nieukorzenione (Rys. 3). Zwykle mamy do czynienia z tymi pierwszym, a różnica między nimi polega na wybraniu w sposób arbitralny tzw. korzenia. Korzeń jest sztucznie utworzonym węzłem i prowadzi bezpośrednio do grupy zewnętrznej czyli najbardziej oddalonej ewolucyjnie od innych sekwencji. Lokalizację korzenia można wyznaczyć opierając się na zasadzie zegara molekularnego lub przez dodanie sekwencji pochodzącej od organizmu odleglejszego ewolucyjnie od każdej analizowanej sekwencji. W obu

przypadkach zlokalizowanie korzenia nie jest pewne, ponieważ w pierwszym przypadku zakłada się poprawność działania zegara molekularnego, co oznacza, że nowo powstałe gałęzie ewoluują w takim samym tempie. W drugim przypadku, aby dodać grupę zewnętrzną musimy posiadać pewien poziom wiedzy dotyczący analizowanych sekwencji (Felsenstein, 2004). Czasem informacji takich po prostu nie ma. Można także próbować wyciągać wnioski analizując posiadane dane. Jednym z możliwych rozwiązań może być wykorzystanie niestacjonarnych modeli substytucji w połączeniu z metodami największego prawdopodobieństwa (Yap i Speed, 2005).



Rys. 4 Drzewo obrazujące stosunki filogenetyczne panujące u naczelnych. a) drzewo nieukorzenione; b) dwa drzewa ukorzenione mają jednakową topologię, a pozorne różnice wynikają jedynie z przestawienia (obrotu) wokół zaznaczonych węzłów. Drzewa powstały w oparciu o kompletne sekwencje mitochondrialne analizowanych organizmów pobrane z GenBank.

Prócz korzenia w drzewie wyróżniamy gałęzie, węzły zewnętrzne (liście), którymi najczęściej są analizowane sekwencje oraz węzły wewnętrzne będące hipotetycznymi sekwencjami pośrednimi między sekwencjami znajdującymi się powyżej miejsca lokalizacji rozpatrywanego węzła (Rys. 3). Sposób ułożenia poszczególnych gałęzi nazywamy topologią drzewa. Analizując kilka sekwencji należy rozważyć wszystkie możliwe układy, które mogą powstać, uwzględniając, że przestawienie dowolnego węzła nie oznacza zmiany topologii drzewa (Rys. 4).
Ogólnie liczba różnych topologi drzew nieukorzenionych dla N sekwencji wynosi:

$$B_{ur}(N) = \prod_{N=3} (2N - 5)$$

(Strimmer, 1997)

Zaś liczba drzew ukorzenionych wynosi odpowiednio:

$$B_r(N) = \prod_{N=2} (2N - 3)$$

Jak widać liczba drzew ukorzenionych dla N sekwencji jest o jeden wykładnik wyższa od liczby drzew nieukorzenionych. Z równań wynika, że ilość możliwych topologii rośnie wykładniczo, tak, że nawet dla względnie małej liczby sekwencji (N) ich ilość szybko dąży do nieskończoności (Tabela2). Oczywiste jest więc, że badanie pokrewieństwa wielu taksonów wymaga sprawdzenia olbrzymiej ilości możliwych drzew i dla większej liczby sekwencji byłoby zbyt czasochłonne, a wręcz niemożliwe. W związku z tym metody wyszukiwania drzew polegają na zastosowaniu algorytmów heurystycznych, które sprawdzają jedynie niewielką część możliwych topologii dzięki przyjęciu pewnych założeń, które mają za zdanie odrzucić błędne topologie jeszcze zanim zostaną wykonane główne obliczenia.

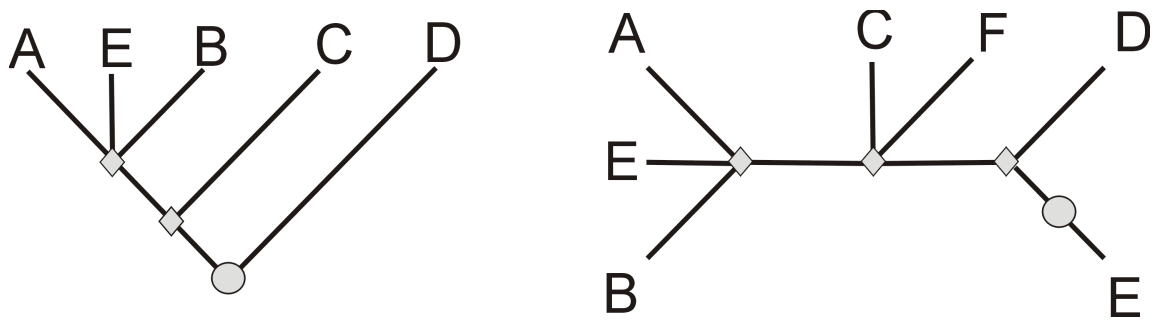
W ogólnym założeniu dąży się do pełnego zobrazowania relacji między sekwencjami czego konsekwencją ma być drzewo o strukturze binarnej czyli drzewo posiadające jeden węzeł stopnia drugiego (z dwoma wychodzącymi od niego gałęziami; węzeł ten nazywamy korzeniem), zaś pozostałe węzły stopnia pierwszego (liście) lub trzeciego (węzły wewnętrzne). Czasem jednak dane nie pozwalają osiągnąć takiego układu. W takim przypadku drzewo zawiera węzły o stopniu

wyższym niż 3 (Rys. 5). Może to wynikać z różnych przyczyn. Drzewa o nie w pełni rozwiązanej strukturze mogą być wynikiem naturalnie panujących stosunków między analizowanymi genami, gdzie jednocześnie powstają więcej niż dwa geny (politomia twarda) lub są błędem wynikającym z zastosowanych danych (nieodpowiednie lub/i za krótkie sekwencje; politomia miękka). Ogólnie przyjmuje się, że w większości przypadków mamy do czynienia z politomią miękką (Cotton, 2003).

Tabela 2. Zależność liczby możliwych drzew od liczby analizowanych cech (sekwencji, organizmów itp.).

N (liczba badanych sekwencji)		
drzewa ukorzenione	drzewa nieukorzenione	liczba możliwych drzew
2	3	1
3	4	3
4	5	15
5	6	105
6	7	945
:	:	:
10	11	34 459 425
11	12	654 729 075
:	:	:
:	:	:
20	21	8 200 794 532 637 891 559 375
$\prod_{N=2} (2N-3)$	$\prod_{N=3} (2N-5)$	

Oprócz graficznej reprezentacji drzew można je zapisać w zwięzłej formie przy użyciu nawiasów. Taki zapis jest szczególnie korzystny w czasie porównywania odległości między poszczególnymi topologiami, która dla binarnych drzew nieukorzenionych równa się podwojonej liczbie węzłów wewnętrznych którymi różnią się analizowane drzewa. Przykładowo drzewo ukorzenione zilustrowane na rysunku 5 można zapisać jako (D (C (B (A, E, B)))), zaś drzewo nieukorzenione jako ((A, E, B) ((C, F) (D, F))) (Nei i Kumar, 2000).



Rys. 5 Politomia drzew ukorzenionego i nieukorzenionego. Oznaczenia jak na rysunku 3.

2.2 CZYNNIKI WPŁYWAJĄCE NA EWOLUCJĘ SEKWENCJI NUKLEOTYDOWYCH

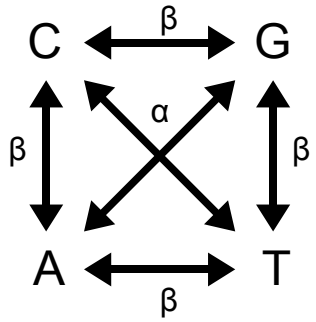
Pierwszym etapem budowy drzewa jest wykonanie zestawienia analizowanych sekwencji (multiple sequence alignment, MSA), które ma za zadanie przypisanie poszczególnym pozycją jednej sekwencji odpowiadającym im pozycją w innych sekwencjach. Dopasowanie sekwencji wymaga wstawienia przerw tak, aby otrzymać jak najlepszy wynik. Istnieje cała gama programów wykonujących ten etap analizy jednak najczęściej używanym jest ClustalX i ClustalW.

Najprostszym współczynnikiem określającym stopień różnicy między dwoma sekwencjami jest stosunek (p) nukleotydów, którymi różnią się dwie sekwencję (n_d) do całkowitej liczby nukleotydów (n).

$$p = n_d/n$$

Współczynnik ten nazywany jest także odległością p między sekwencjami. Odległość p poza prostotą niewiele ma do zaoferowania, ponieważ jest współczynnikiem skrajnie upraszczającym rzeczywistość. Nie uwzględnia on substytucji wstecznych (np. A→C→A) ani innych dostępnych informacji wynikających z charakteru zasad lub aminokwasów. Tą pierwszą wadę da się wyeliminować stosując odległość uwzględniającą rozkład Poissona (odległość PC) lub rozkład gamma. Jeśli chodzi o drugie zastrzeżenie to od dawna wiadomo, że proporcje poszczególnych zasad nie są jednakowe i w zależności od organizmu bądź sekwencji mogą być dalekie od tych które można by oczekiwać z praw statystyki. Zawartość zasad GC może się wahać od 30% do ponad 60%.

Podobnie substytucje poszczególnych zasad nie są jednakowe. Generalnie można wyróżnić dwa rodzaje substytucji, mianowicie transwersje i tranzycje. Transwersje są podstawieniami jednej puryny (adenina lub guanina) w drugą purynę lub pirymidyny (tymina, cytozyna) w inną pirymidynę, a transwersje to zamiana puryny w pirymidynę lub na odwrót (Rys. 6).



Rys. 6 Typy substytucji. α – tranzycja (A↔G, C↔T); β- transwersja (pozostałe).

Współczynnik tranzycji do tranwersji (R) dla większości genów jądrowych mieści się w przedziale 0.5-2.0, ale dla DNA mitochondrialnego może dochodzić aż do 15. Wynika to z prostej przyczyny, według której tranzycje nukleotydów są bardziej prawdopodobne, ponieważ zamiana zachodzi między związkami o podobnej budowie, natomiast transwersje dotyczą odmiennych strukturalnie związków. Następnym czynnikiem, który należy wziąć pod uwagę jest różne tempo ewolucji między poszczególnymi miejscami w obrębie kodonów. Wynika to w prosty sposób z degeneracji kodu genetycznego, który, aby jeszcze bardziej skomplikować sytuację nie jest uniwersalny dla wszystkich organizmów. Przyglądając się budowie kodu genetycznego łatwo zauważyć, że pewne zmiany zachodzą łatwiej niż inne, ponieważ nie powodują zmiany odczytu kodu. Zmiany takie nazywamy synomicznymi i dotyczą praktycznie wszystkich zmian trzeciego nukleotydu kodonów i niektórych zmian zachodzących na pierwszej pozycji. Przykładowo w kodonie CUA kodującym leucynę mutacja A na dowolny nukleotyd np. U nie spowoduje żadnej zmiany w odczycie kodonu, CUU nadal koduje leucynę. Podobnie substytucja C→U pozostaje zmianą cichą. Dzięki temu zmiany synomiczne nie będą wpływać na budowę białka, a więc pozostaną w dużym stopniu neutralne. Zupełnie inaczej przedstawia się sprawa zmian niesynomicznych czyli takich, które powodują zmianę odczytu kodu.

W tym przypadku ich obecność jest widoczna na poziomie białka i może znacząco wpływać na jego budowę i funkcję. Oczywiście większość tych zmian będzie szkodliwa i dlatego zmiany niesynomiczne będą szybko eliminowane i jedynie nieliczne zostaną na stałe włączone do genomu i się utrzymają (jest to tzw. selekcja oczyszczająca). Substytucję niesynomiczne dotyczą niektórych zmian pierwszej pozycji kodonów i wszystkich zmian w obrębie drugiej pozycji, które w tym ostatnim przypadku zawsze skutkują zmianami odczytywanego aminokwasu lub mutacjami nonsensownymi. Wracając do naszego przykładu zmiana U w obrębie CUA na inny nukleotyd np. G spowoduje powstanie kodonu kodującego zupełnie inny aminokwas (w tym przypadku argininę).

Kolejnym ważnym czynnikiem jest dostępność poszczególnych tRNA dla odpowiednich kodonów. To, że dany aminokwas może być kodowany przez np. cztery różne kodony wcale nie oznacza, że tRNA komplementarne do nich będą występować w komórce w takiej samej ilości. Wręcz przeciwnie, w większości przypadków jeden z tRNA występuje obficie, a reszta w znikomych ilościach. W ten sposób kodon dla którego tRNA występuje w dużych ilościach będzie preferowany. Zjawisko to widać szczególnie w odniesieniu do genów podlegających intensywnej ekspresji. Przykładowo częstość kodonów kodujących glicynę w genach polimerazy RNA *Escherischia coli* wskazuje, że o ile kodony GGU i GGC stanowią zdecydowaną większość to brak całkowicie kodonów GGA i GGG. Współczynnikiem opisującym to zjawisko jest względny poziom użycia kodonów synomicznych (relative synonymous codon usage, RSCU) odpowiadający stosunkowi obserwowanej częstości występowania kodonu do częstości oczekiwanej przy założeniu jednakowego wykorzystania poszczególnych kodonów (Nei i Kumar, 2000).

2.3 MODELE EWOLUCJI SEKWENCJI NUKLEOTYDOWYCH

Istnieje wiele modeli opisujących sposób zachodzenia substytucji nukleotydowych. Pierwszym i zarazem najprostszym modelem jest model

zaprezentowany w 1969 roku przez Jukesa i Cantora (JC69). Zakłada on, że substytucje zachodzą w każdym miejscu z tą samą częstotliwością i każdy nukleotyd może ulec zmianie na dowolny inny ze stałą częstotliwością α. Jak zaznaczono wcześniej ilość traznzycji jest zwykle większa niż transwersji co może znacznie wpływać na otrzymany wynik. W związku z tym zaproponowano inny model (K2P) w którym częstość tranzycji określa parametr α, a częstość transwersji parametr β. W modelu K2P częstość wszystkich substytucji wynosi α + 2β. Kolejnym modelem jest jest model zaproponowany przez Felsensteina (F81), który bierze pod uwagę częstość poszczególnych nukleotydów. Modelem dodatkowo uwzględniającym różną zawartość zasad GC jest model Tamury (T92). Inny często używany model HKY jest połączeniem modelu K2P i F81 i uwzględnia zawartość GC i stosunek transzycji do transwersji. Modelem najbardziej złożonym uwzględniającym 10 oddzielnych parametrów dla każdej możliwej substytucji jest model GTR w którym jedynym przyjmowanym założeniem jest odwracalność ewolucji (Tabela 3). Należy zaznaczyć, że o ile bardziej złożone metody wykorzystujące większą ilość parametrów z założenia powinny lepiej obrazować odległość między sekwencjami o tyle wariancja d wzrasta wzrasta proporcjonalnie do liczby parametrów, tak, że w niektórych przypadkach użycie prostszego modelu może dać taki sam rezultat jak wykorzystanie skomplikowanego modelu wymagającego dużych nakładów mocy obliczeniowej (Nei i Kumar, 2000). Omówione pokrótce powyżej modele substytucji zaliczane są do jednej klasy modeli procesów odwracalnych w czasie (REV), które zakładają możliwość wnioskowania o ewolucji w oparciu o założenie jej odwracalności. Jednak podejście takie niekoniecznie musi być prawdziwe w związku z czym coraz częściej stosuje się inne bardziej wyrafinowane modele zakładające nieodwracalność procesów substytucji. Mimo, że metody te powstały niedawno dają one obiecujące wyniki i już teraz wykorzystywane są do umiejscowienia korzenia w oparciu o analizowane dane, a nie przez subiektywne wybranie grupy zewnętrznej przez badacza (Yap i Speed, 2005).

Tabela 3 Modele substytucji nukleotydów (według Nei i Kumar, 2000; zmienione).

	A	**T**	**C**	**G**	**A**	**T**	**C**	**G**
		model JC69				model T92		
A	–	α	α	α	–	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$
T	α	–	α	α	$\beta\theta_2$	–	$\alpha\theta_1$	$\beta\theta_1$
C	α	α	–	α	$\beta\theta_2$	$\alpha\theta_2$	–	$\beta\theta_1$
G	α	α	α	–	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	–
		model K2P				model HKY		
A	–	β	β	α	–	βg_T	βg_C	αg_G
T	β	–	α	β	βg_A	–	αg_C	βg_G
C	β	α	–	β	βg_A	αg_T	–	βg_G
G	α	β	β	–	αg_A	βg_T	βg_C	–
		model F81				model GTR		
A	–	αg_T	αg_C	αg_G	–	ag_T	bg_C	cg_C
T	αg_A	–	αg_C	αg_G	ag_A	–	dg_C	eg_C
C	αg_A	αg_T	–	αg_G	bg_A	dg_T	–	fg_C
G	αg_A	αg_T	αg_C	–	cg_A	eg_T	fg_C	–

Oznaczenia: g_A, g_T, g_C i g_G oznaczają częstość nukleotydów; $\theta_1 = g_G + g_C$; $\theta_2 = g_A + g_T$;

2.4 CZYNNIKI WPŁYWAJĄCE NA EWOLUCJĘ BIAŁEK

Także na poziomie białka istnieje szereg czynników wpływających na ich ewolucję, a także na ewolucję DNA kodującego dane białko. Ze względów wspomnianych wcześniej (patrz rozdz. 2.2) wynika, że częstość zmian jednego aminokwasu w inny nie jest czysto losowa. Aminokwasy o właściwościach podobnych będą tutaj wyraźnie preferowane. W praktyce oznacza to, że istnieje większe prawdopodobieństwo, że przykładowo leucyna ulegnie zmianie w izoleucynę niż w argininę. Aspekt ten wynika ze wszystkich właściwości danego aminokwasu takich jak hydrofobowość (regiony hydrofobowe pogrążone zwykle wewnątrz struktury białka, gdzie oddziałują z innymi aminokwasami niemal z każdej strony, wolniej ewoluują niż aminokwasy hydrofilowe swobodnie wystające do środowiska), kwasowość czy jego wielkość. Ponadto szybkość ewolucji w obrębie

domen białka będzie różna i ogólnie jest ona odwrotnie skorelowana z znaczeniem danego aminokwasu w budowie struktur wyższego rzędu czy centrów aktywnych enzymów. Wykorzystanie tych informacji może korzystnie wpłynąć na poprawność analizy filogenetycznej.

Tabela 4 Macierz PAM250 (wartości na głównej przekątnej i pod nią) i macierz BLOSUM62 (wartości powyżej głównej przekątnej, wartości dla przekątnej podano poniżej w wierszu oznaczonym *; na podstawie Baxevalis i Ouellette, 2004).

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
A	2	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-2	6	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	1	0	-4	-2	-3
N	0	0	2	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	0	-1	2	4	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	-2	-4	-4	-5	12	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	0	1	1	2	-5	4	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	0	-1	1	3	-5	2	4	-2	0	-3	-3	1	-2	-2	-1	0	-1	-3	-2	-2
G	1	-3	0	1	-3	-1	0	5	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-1	2	2	1	-3	3	1	-2	6	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-1	-1	-1	-1	1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-4	-2	-2	1	3	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	-1	-1	-4	-3	-2
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-3	-2	-2
T	1	-1	0	0	-2	-1	0	-1	0	0	-2	0	-1	-3	0	1	3	-2	-2	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	2	-3
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-1
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4
*	4	5	6	6	9	5	5	6	8	4	4	5	5	6	7	4	5	11	7	4

Podobnie jak w przypadku sekwencji nukleotydowych podobieństwo między dwoma sekwencjami możemy opisać za pomocą liczby oczekiwanych substytucji d, którą dla większej wiarygodności możemy skorygować używając rozkładów gamma i PC. Wartość d dla białek opisuje się także przy pomocy odpowiednio

zmodyfikowanych modeli JC69 i K2P, jednak ich użyteczność jako bardzo ogólnych jest wątpliwa, obecnie stosuje się coraz częściej odległość *Scoredist*, która zakłada poprawkę logarytmiczną obserwowanej wartości d w oparciu o macierz BLOSUM62 (Sonnhammer i Hollich, 2005).

2.5 MODELE EWOLUCJI BIAŁEK

Ewolucję białek próbuje odtworzyć się poprzez zastosowanie odpowiednich macierzy substytucji o rozmiarze 20 x 20 na podstawie danych empirycznych. Najpierw pojawiły się macierze oparte na modelu akceptowanych mutacji punktowych (PAM) w których jedna jednostka PAM odpowiada różnicy ewolucyjnej między dwoma sekwencjami wynoszącej 1%. Model ten uwzględnia zajście wstecznych substytucji i występowanie pewnych preferencji do częstszych substytucji jednych aminokwasów od innych. W ten sposób możliwe jest wyznaczenie macierzy PAM wyższych rzędów (PAM200-300). Podstawowa macierz o wartości 1 PAM została wyznaczona przez Dayhoff w oparciu o blisko spokrewnione sekwencje, a następnie otrzymane wyniki ekstrapolowano do innych odległości PAM (Tabela 4). Macierze PAM o niskim stopniu są wykorzystywane do analizy blisko spokrewnionych sekwencji, zaś te o większym stopniu do analizy sekwencji odleglejszych ewolucyjnie. Innym powszechnie stosowanym typem macierzy są macierze BLOSUM (Henikoff i Henikoff, 1992) oparte o dane zawarte w bazie danych BLOCKS (Henikoff i Henikoff, 1991). W przeciwieństwie do macierzy PAM macierze BLOSUM powstały w oparciu o sekwencje oddalone ewolucyjnie. Ich liczba oznacza poziom homologi sekwencji, które zostały użyte do stworzenia danej macierzy. Przykładowo, najczęściej używana macierz BLOSUM62 powstała w oparciu o sekwencje wykazujące co najmniej 62% identyczności. Pomiędzy poszczególnymi macierzami BLOSUM nie ma żadnej zależności matematycznej tak jak to miało miejsce w przypadku macierzy PAM. Oprócz tych macierzy powstał cały szereg macierzy wyznaczonych w oparciu o sekwencje białek mitochondrialnych (mtREV24) lub inne bazy danych (JTT, VT, WAG; Whelan i Goldman, 2001).

Ponadto dostępne są specjalnie przygotowane macierze pod kątem rybosomalnego RNA (bactSLT, eukSLT, euk23SLT, mitoSLT; Smith i wsp., 2004).

2.6 METODY TWORZENIA DRZEW

Istnieje wiele metod statystycznych wykorzystywanych do konstrukcji drzew. Generalnie metody te można podzielić na kilka grup: metody oparte na odległości (distance methods), metoda parsymonii (największej oszczędności, MP), metoda największej wiarygodności (ML) i metody Bayesa. Oprócz tych metod o potwierdzonej skuteczności próbuje się wykorzystać wiele innych, często z pozytywnym skutkiem. Jedną z najbardziej obiecujących możliwości jest wykorzystanie programowania genetycznego (Lemmon i Milinkovitch, 2002). Ze względów praktycznych omówione zostaną jedynie wybrane metody skupając się głownie na ich zaletach i wadach, bardziej szczegółowe dane razem z podstawami matematycznymi można znaleść w innych źródłach (Nei i Kumar, 2000; Felsenstein 2004).

Metody oparte na odległości polegają na przedstawieniu różnic między poszczególnymi sekwencjami w postaci liczb, które określają ich wzajemną odlagłość. Zaletą tej grupy metod jest prostota i szybkość, jednak metody te nie są w stanie zobrazować procesu w sposób idealny, ponieważ nie wszystkie zdarzenia ewolucyjne można odtworzyć na podstawie rozpatrywanych danych. Przykładem takich zmian są wielokrotne substytucje wsteczne. W większości przypadków można to zjawisko w pewien sposób zminimalizować wprowadzając poprawkę proporcjonalną do dywergencji sekwencji. Inną wadą tych metod jest to, że w kolejnych etapach analizy wszelkie obliczenia oparte są na podstawie początkowo wyliczonych wartości co może oznaczać utratę ważnych informacji. W zależności od implementacji wyróżnia się tu szereg metod takich jak metoda nieważonych średnich połączeń (UPGMA), metoda ostatnich kwadratów (LS), metoda Fitch-Margoliach (FM), matoda minimalnej ewolucji (ME), metoda przyłączania sąsiada (NJ) i jej modyfikacje (BIONJ). Metoda UPGMA polega na łączeniu gałęzi o największym

podobieństwie według średnich łączonych par. Podejście takie zakłada istnienie zegara biologicznego i dobrze oddaje relacje między analizowanymi sekwencjami tylko, gdy ich tempo substytucji jest względnie stałe. Warunek ten jest rzadko spełniony i dlatego metoda ta jest rzadko stosowana. Inną metodą jest metoda FM polegająca na zmaksymalizowaniu dopasowania między parami sekwencji przez zminimalizowanie odchylenia kwadratowego obserwowanych odległości w stosunku do wszystkich możliwych długości gałezi drzewa. Wadą tej metody jest brak możliwości oceny uzyskanej topologii drzewa, którą ustala się w oparciu o długość gałęzi drzewa (jednak poszczególne wyniki nie są od siebie w pełni niezależne). Ponadto metoda ta jest tym skuteczniejsza im dłuższych sekwencji użyjemy co znacznie ogranicza jej wykorzystanie (Nei i Kumar, 2000). Podobne podejście zakłada metoda ME w której sumuje się długości wszystkich gałęzi każdego z możliwych drzew i wybiera się drzewo o najmniejszej ich wartości. Wadą takiego podejścia jest jego czasochłonność przy dużej liczbie sekwencji czego można w pewnym stopniu uniknąć ograniczając liczbę analizowanych topologii, mimo to musimy przeanalizować (N-1)!/2 różnych drzew (jako punkt startowy podaje się drzewo ustalone za pomocą metody NJ; Kumar, 1996). Jedną z najszybszych metod tej grupy jest NJ będąca swego rodzaju uproszczeniem metody ME. Swoją szybkość zawdzięcza temu, że analiza podlega na sprawdzeniu wybranej grupy topologii. Odbywa się to według następującego algorytmu. Drzewem wyjściowym jest drzewo w kształcie gwiazdy, następnie brane są pod uwagę dwa taksony (potencjalni sąsiedzi czyli sekwencje połączone wspólnym węzłem) i porównuje z innymi. Dwa najbardziej podobne taksony (posiadające najkrótszą długość gałęzi) są wybierane, a między nimi wstawiana jest dodatkowa gałąź łącząca je z pozostałymi taksonami drzewa w kształcie gwiazdy. Postępuje się tak, aż do wyczerpania taksonów (Saitou i Nei, 1987). Wadą tej metody jest to, że uzyskujemy jedno drzewo, które niekoniecznie może być jedynym możliwym rozwiązaniem o ustalonej długości gałęzi.

Metoda parsymonii czyli oszczędności zakłada, że najlepszym rozwiązaniem jest to najprostsze czyli takie drzewo, które wymaga najmniej zmian (substytucji).

Analizowane są jedynie miejsca w których sekwencje się różnią, pozostałe pozycje są usuwane i nie są dalej wykorzystywane. Dodatkowo miejsca takie muszą się różnić przynajmniej w dwóch sekwencjach. Założenie to ogranicza w znacznym stopniu zastosowanie tej metody. Z powodu występowania wstecznych substytucji przy wysokim poziomie dywergencji sekwencji metoda ta nie jest w stanie określić prawidłowej topologii drzewa. Ogólnie metodę tą można stosować gdy, sekwencje mają niski poziom dywergencji ($d \leq 0.1$), tempo substytucji jest podobne w obrębie różnych gałęzi i długość sekwencji jest odpowiednio duża. Pewną poprawę może wnieść zastosowanie parsymonii ważonej w której tranzycjom i tranzwersjom nadawane są inne wagi (Nei i Kumar, 2000). Zaletą i jednocześnie wadą tej metody jest tworzenie wielu optymalnych drzew, czasem ich liczba jest tak duża, że nie pozwala właściwie określić poprawnej topologii (Hedges i wsp., 1991).

Trzecią grupę metod stanowią metody największego prawdopodobieństwa. Generalnie polegają one na znalezieniu takiego modelu ewolucji, który w najlepszy sposób tłumaczy powstanie określonego drzewa w oparciu o analizowane dane. Najpierw obliczane są cząstkowe wartości wiarygodności dla poszczególnych miejsc, które się następnie wymnaża. Zwykle otrzymana wartość jest bardzo małą niską liczbą i dlatego przedstawia się ją w postaci ujemnego logarytmu. Procedurę obliczania prawdopodobieństw należy powtórzyć dla wszystkich możliwych topologii. Kolejnym pracochłonnym etapem jest obliczenie długości gałęzi. Z tych względów metoda ta jest niezwykle wymagająca pod względem mocy obliczeniowej (Sullivan, 2005). Mimo to metodę największej wiarygodności uważa się obecnie za najskuteczniejszą i jest ona wyraźnie preferowana (Gadagkar i Kumar, 2005; Holder, 2001; Piontkivska, 2004). Niezbędny czas obliczeń można zredukować na dwa sposoby. Pierwszy polega na wyjściu od najbardziej skomplikowanego modelu (o największej liczbie parametrów) i jego stopniowym upraszczaniu aż do momentu, gdy jego dalsze upraszczanie wywoła istotną statystycznie różnicę. W ten sposób w zależności od danych z modelu o 10 parametrach można dojść do innego prostszego modelu co znacznie uprości obliczenia i w efekcie skróci czas analizy (Sullivan, 2005). Innym podejściem może być zastosowanie różnych metod heurystycznych,

które wybiorą jedynie nieliczne, najbardziej prawdopodobne topologie i w oparciu o nie wykonają niezbędne obliczenia. Istnieje szereg różnych algorytmów wykonujących to zadanie (dla przykładu patrz Strimmer, 1977) jednak ich omówienie wykracza poza ramy niniejszej pracy. Szczególnie to drugie podejście będzie wykorzystywane przez mnie w praktycznej części pracy.

2.7 WIARYGODNOŚĆ OTRZYMANYCH WYNIKÓW

Standardową procedurą sprawdzającą prawidłowość otrzymanej topologii drzewa jest metoda bootstrap. Polega ona na wygenerowaniu określonej liczby (zwykle 1000) zestawów sekwencji powstałych z sekwencji użytych do budowy drzewa przez ich losowy wybór ze zwracaniem spośród zasad lub aminokwasów, które wchodzą w skład danej sekwencji. Następnie całą analizę powtarza się odpowiednią liczbę razy. Otrzymane wyniki porównywane są z pierwotnym drzewem przez co otrzymujemy procentowe poparcie poprawności danego rozgałęzienia (Nei i Kumar, 2000). Mimo, że metoda ta jest niezwykle użyteczna to jednak powoduje wielokrotne przedłużenie czasu analizy, czasem wręcz uniemożliwiając jej wykonanie. Pewnym rozwiązaniem tego problemu może być użycie innych metod. Jedną z nich jest mapowanie prawdopodobieństwa za pomocą kwartetów, które sprawdza czy dany zbiór sekwencji jest odpowiedni do analizy (Strimmer, 1997). Metoda ta polega na sprawdzaniu stosunków fiogenetycznych między losowo wybranymi czterema sekwencjami (kwartetami). Przy odpowiednio licznej próbie na tej podstawie budowane jest uśrednione drzewo końcowe. W przypadku czterech sekwencji istnieją trzy możliwości zbudowania drzewa nieukorzenionego. W zależności od danych drzewo takie może posiadać jednoznaczną (regiony A_1, A_2, A_3), częściowo rozwiązaną (regiony A_{12}, A_{23}, A_{13}), bądź nierozwiązaną strukturę (region A_{123}). Ich procentowy udział pozwala z góry przewidzieć czy analizowane dane są odpowiednie do budowy drzewa. Całość można nanieść na diagram w kształcie trójkąta (Rysunek 12a). Duży udział kwartetów w regionie A_{123} oznacza, że drzewo wynikowe będzie wykazywać silną politomię.

3. MATERIAŁY I METODY

Do badań wykorzystano nienadmiarowe sekwencje kręgowców zamieszczone w Histone Sequence Datebase umieszczone w NHGRI/NCBI (stan na maj 2005 rok, Sullivan i wsp., 2002) z pominienciem nielicznych sekwencji, które były zbyt krótkie. Ponadto przeszukano zasoby GenBank i dołączono szereg sekwencji przewidzianych drogą automatyczną. Dodatkowo do ukorzenienia drzew jako grupy zewnętrznej użyto sekwencji histonowych jeżowców *Lytechinus pictus*, *Parechinus angulosus* i *Strongylocentrotus purpuratus.* W sumie analizą objęto 99 sekwencji białkowych i 83 sekwencji nukleotydowych z 24 gatunków kręgowców i 3 wyżej wymienionych bezkręgowców. Geny histonów poddano obróbce (analizę ograniczono do regionów kodujących białko). Szczegółowe dane na temat numerów dostępu GenBank wykorzystanych sekwencji i ich długość zamieszczono w Tabeli 8. Dopasowanie sekwencji MSA przeprowadzono za pomocą programu ClustalX i ClustalW (Thompson i wsp., 2003). W przypadku nukleotydów uzyskano MSA o długości 907 nukleotydów (Rysunek 7), a dla białek MSA o długości 308 aminokwasów (Rysunek 8).

Do budowy drzew metodami dystansu wykorzystano program MEGA3.1 (Kumar i wsp., 2004). Ze względu na złożoność obliczeń analizy metodą ML przeprowadzono za pomocą programów heurystycznych TREEFINDER (Jobb i wsp., 2004) i TREE-PUZZLE v.5.2 (Strimmer i von Haeseler, 1996), a nie z wykorzystaniem pakietu PHYLIP czy programu PAUP. Wiarygodność uzyskanych wyników sprawdzono przy użyciu metody bootstrap.

Do analizy sekwencji białkowych matodami ML zastosowano model Dayhoff (PAM) uwzględniając rozkład gamma (dyskretyzacja na 6 przedziałów) dla miejsc zmiennych i miejsc konserwatywnych (niezmiennych) ewolucyjnie (Dayhoff+G+I). Model ten został wybrany na podstawie porównania wartości największego prawdopodobieństwa (-lnL), kryterium informacyjnego Akaike (AIC, AICc) i kryterium informacyjnego Bayesa (BIC) określonych przy pomocy programu Prottest (Abascal i wsp., 2005). Początkowe drzewo ustalono metodą BIONJ (Tabela 5).

Analizę sekwencji nukleotydowych przeprowadzono w oparciu o model HKY+G+I, który wybrano na podstawie wartości -lnL, AIC uzyskanych za pomocą serwera online FindModel opartego na programie Modeltest v.3.7 (Posada i Crandall, 1998). Drzewo początkowe ustalono za pomocą metody NJ według modelu ważonych odległości JC.

Tabela 6. Optymalizacja modelu ewolucji białek histonowych według kryteiów AIC, AICc i BIC. W nawiasach podano pozycje spośród 20 przeanalizowanych modeli (część danych pominięto).

Model	AIC	AICc-1	AICc-2	AICc-3	BIC-1	BIC-2	BIC-3
Dayhoff+I+G	**0.97(1)**	0.03(2)	**0.78(1)**	**0.96(1)**	**0.81(1)**	**0.78(1)**	**0.42(2)**
Dayhoff+G	0.03(2)	**0.97(1)**	0.22(2)	0.04(2)	0.19(2)	0.22(2)	0.58(1)
WAG +I+G	0.00(3)	0.00(4)	0.00(4)	0.00(3)	0.00(4)	0.00(4)	0.00(4)
VT+I+G	0.00(5)	0.00(7)	0.00(5)	0.00(5)	0.00(5)	0.00(5)	0.00(6)
JTT+I+G	0.00(6)	0.00(8)	0.00(7)	0.00(6)	0.00(7)	0.00(7)	0.00(8)
Blosum62+I+G	0.00(17)	0.00(18)	0.00(18)	0.00(17)	0.00(18)	0.00(18)	0.00(18)

AIC: kryterium informacyjne Akaike
AICc-x: kryterium informacyjne Akaike drugiego rzędu
BIC-x: kryterium informacyjnego Bayesa
AICc/BIC-1: rozmiar próby (długość sekwencji) (308.0)
AICc/BIC-2: rozmiar próby (suma entropii Shanona poszczególnych miejsc do całkowitego zestawienia) (459.6)
AICc/BIC-3: rozmiar próby (długość sekwencji x liczba sekwencji x uśredniona (0-1) entropia Shanona) (10528.4)

Tabela 5. Alfabetyczny spis sekwencji białkowych i nukleotydowych

Gatunek	Wariant histonu	Oznaczenie białka	Kod dostępu białka	Dłu-gość	Oznaczenie sekwencji nukleotydowej	Kod dostępu GenBank	Długość	
Anas platyrhynchos	-	Ap	CAA29495	218	Ap	X06128	1941 (654)	&
Anguilla japonica	-	Aj	BAB91236	145	Aj	AB073744	472 (435)	
Anser anser anser	H5	Aa5	HSGS5 P02258	193 193				
Bos taurus	H1.2 (H1d)	Bt12	XP_591742	213	Bt12	XM_591742	1141 (639)	*
Bos taurus	H1.0	Bt10	XP_583447	194	Bt10	XM_583447	1844 (582)	*
Bufo bufo gagarizans	-	Bb	AAK66966	224	Bb	AF255740	839 (672)	
Cairina moschata	H1	Cm	S01262	218				
Cairina moschata	H5	Cm5	P06513 CAA25530	194	Cm5	X01065	1175 (582)	
Canis_familis	$H1^0$	Cf	XP_852794	194	Cf	XM_847701	867 (582)	*
Carassius auratus	-	Ca	AAO72080	191	Ca	AY184811	584 (573)	
Danio rerio	-	Dr	AAH76144	163	Dr	BC076144	543 (489)	&
Danio rerio	-	Dr_a	XP_698497	199	Dr_a	XM_693405	619 (597)	*
Danio rerio	-	Dr_b	XP_697886	200	Dr_b	XM_692794	603 (600)	*
Danio rerio	H1M	Dr1Ma	AAM22974	257	Dr1M_a	AF499607	917 (771)	*
Danio rerio	H1M	Dr1Mb	NP_898894	257	Dr1M_b	NM_183071	1374 (771)	*

Tabela 5. Alfabetyczny spis sekwencji białkowych i nukleotydowych

Gatunek	Wariant histonu	Oznaczenie białka	Kod dostępu białka	Dłu-gość	Oznaczenie sekwencji nukleotydowej	Kod dostępu GenBank	Długość	
Danio rerio	H1.5	Dr15	XP_688137	176	Dr15	XM_683045	644 (528)	*
Danio rerio	H1°	Dr10	NP_955846	199	Dr10	NM_199552	1365 (597)	*
Danio rerio	H1X	Dr1X	NP_954970	192	Dr1X	NM_199276	1351 (576)	*
Gallus gallus	H1.2 (H1d)	Gg12	NP_001035732 XP_416194	219 219	Gg12	NM_001040642 XM_416194	660 1531	
Gallus gallus	H1.02	Gg102	XP_425456	218	Gg102	XM_425456	657 (654)	*
Gallus gallus	H1.10	Gg110	XP_416189	220	Gg110	XP_416189	697 (660)	*
Gallus gallus	-	Gg	P09987	218	Gg	J00863	1098 (654)	
Gallus gallus	H5	Gg5	P02259 70678 AAA48798 CAA24994	190 190 190 190	Gg5	J00870 X00169	1843 (570) 6090	
Gallus gallus	H1.11R	Gg111R_XP	XP_425470	219	Gg111R_XM	XM_425470	660	*
Gallus gallus	H1.11R	Gg111R_P0	P08288 AAA48790 86291	219 219 219	Gg111R_M	M17020 M17020	1264 1264 (657)	
Gallus gallus	H1.10 (H1c)	Gg110_XP	XP_425466 NP_001035733	225 225	Gg110_XM Gg110_NP	XM_425466 NP_001040643	678 770 (675)	* *
Gallus gallus	H1.10	Gg110_A	A28456	220				
Gallus gallus	H1.03	Gg103	P08285 86287 AAA48787	224 224 224	Gg103	M17021 M17021	1140 (672) 1140	
Homo sapiens	H1X	Hs1X	NP_006017	213	Hs1X	NM_006026	1503 (639)	
Homo sapiens	H1t	Hs1t_P	P22492 87670 1708098 AAA35944	207 207 207	Hs1t_M	M60094 87670 M60094 M60094	1759 (621) 1759 1759	
Homo sapiens	H1t	Hs1t_AAN	AAN06701	207	Hs1t_AF	AF531301	1380 (621)	
Homo sapiens	H1t	Hs1t_AAH	AAH69517	221	Hs1t_BC	BC069517	718 (621)	
Homo sapiens	H1t	Hs1t_AAA	AAA19936 NP_005314	207	Hs1t_M9 Hs1t_NM	M97755 NM_005323	874 725 (621)	
Homo sapiens	$H1^0$	Hs10_AAH	AAH29046	194	Hs10_BC	BC029046	2193 (582)	
Homo sapiens	$H1^0$	Hs10_NP	NP_005309	194	Hs10_NM	NM_005318	2336 (582)	
Homo sapiens	H1.2 (H1c)	Hs12	NP_005310	213	Hs12	NM_005319	732 (639)	
Homo sapiens	H1.3 (H1d)	Hs13	NP_005311	221	Hs13	NM_005320	777 (663)	
Homo sapiens	H1.4 (H1e)	Hs14	NP_005312	219	H14	NM_005321	785 (657)	
Homo sapiens	H1.5 (H1b)	Hs15_NP	NP_005313	226	Hs15	NM_005322	790 (678)	
Homo sapiens	H1.1 (H1a)	Hs11	AAN06699	215	Hs11	AF531299	1620 (645)	
Homo sapiens	H1.5 (H1b)	Hs15_12	1208320A	219				
Ictalurus punctatus	-	Ip	AAQ99138	203	Ip	AY324398	956 (609)	&
Lytechinus pictus	-	Lp	CAA28177	210	Lp	X04488	1171 (630)	&
Macaca mulatta	H1t	Rh1t	P40286 7439622 AAA19937	208 208	Rh1t	M97756 M97756	874 (624) 874	
Mus musculus	H1X	Mm1X	NP_941024	188	Mm1X	NM_198622	1126 (564)	
Mus musculus	H1t	Mm1t_NP	NP_034507	208	Mm1t_NM	NM_010377	727 (624)	
Mus musculus	H1t	Mm1t_CAA	CAA51325	208	Mm1t_X	X72805	2808 (624)	
Mus musculus	H1e	Mm1e	NP_056602	219	Mm1e	NM_015787	1954 (657)	

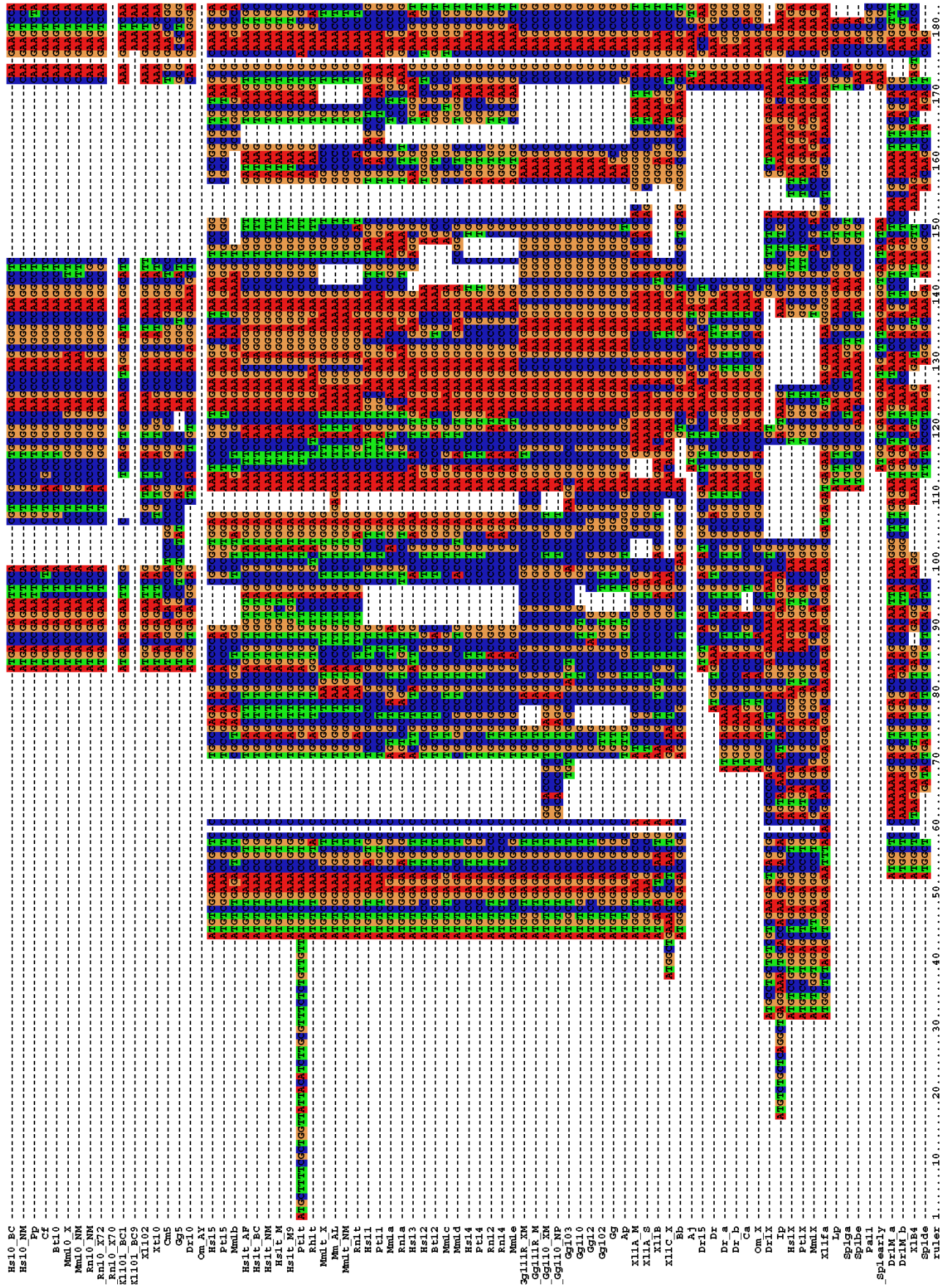
Tabela 5. Alfabetyczny spis sekwencji białkowych i nukleotydowych

Gatunek	Wariant histonu	Oznaczenie białka	Kod dostępu białka	Dłu-gość	Oznaczenie sekwencji nukleotydowej	Kod dostępu GenBank	Długość	
Mus musculus	H1d	Mm1d	NP_663759	221	Mm1d	NM_145713	2975 (663)	
Mus musculus	H1c	Mm1c	NP_056601	212	Mm1c	NM_015786	1563 (636)	
Mus musculus	H1b	Mm1b	NP_064418	223	Mm1b	NM_020034	672 (669)	
Mus musculus	H1a	Mm1a	NP_085112	213	Mm1a	NM_030609	743 (639)	
Mus musculus	$H1^0$	Mm10_CAA	CAA31569	194	Mm10_X	X13171	1354 (582)	
Mus musculus	$H1^0$	Mm10_NP	NP_032223	194	Mm10_NM	NM_008197	2304 (582)	
Mus musculus	-	Mm_CAI	CAI24904	209	Mm_AL	AL592149	627	&
Oncorhynchus mykiss	-	Om_P	P06350 70668 CAB37646	207 206 207	Om_X	X02624 X02624	1728 (621) 1728	
Oncorhynchus mykiss	-	Om_AAN	AAN86579	144	Om_AY	AY150299	577 (432)	
Oryctolagus cuniculus	H1.3	Oc13	P02251 HSRB13	213 213				
Pan troglodytes	H1.4 (H1b)	Pt14	XP_527259	219	Pt14	XP_527259	660	*
Pan troglodytes	H1t	Pt1t	XP_527257	223	Pt1t	XP_527257	672	*
Pan troglodytes	H1.1	Pt11	XP_527252	215	Pt11	XP_527252	648	*
Pan troglodytes	H1X	Pt1X	XP_526304	213	Pt1X	XP_526304	642	*
Pan troglodytes	H1.5 (H1a)	Pt15	XP_527284	226	Pt15	XP_527284	681	*
Parechinus angulosus	-	Pa	Q7M409 103649	206 206				
Parechinus angulosus	H1.1	Pa11	AAA17393	180	Pa11	U07825	540	
Pongo pygmaeus	H1°	Pp	CAI29624	194	Pp	CR925976	2239 (582)	* &
Rattus norvegicus	H1°	Rn10_NP	NP_036710 AAH61842 P43278 631842	194 194 194 177	Rn10_NM	NM_012578 BC061842 NP_036710	1887 1887 1887 (582)	
Rattus norvegicus	H1°	Rn10_CAA9	CAA51199	194	Rn10_X72	X72624	1779 (582)	
Rattus norvegicus	H1°	Rn10_CAA0	CAA50020	164	Rn10_X70	X70685	1711 (492)	
Rattus norvegicus	H1.2 (H1d)	Rn12_P	P15865 92378	219 217				
Rattus norvegicus	H1.2 (H1d)	Rn12_C	CAA47734	219				
Rattus norvegicus	H1.2 (H1d)	Rn12_A	AAA41327	217	Rn12	M31229	1191 (651)	
Rattus norvegicus	H1a	Rn1a	XP_225330	214	Rn1a	XP_225330	693 (642)	*
Rattus norvegicus	H1.4	Rn14	NP_579819	219	Rn14	NM_133285	660 (657)	
Rattus norvegicus	H1t	Rn1t	NP_036711	208	Rn1t	NP_036711	745 (624)	
Salmo trutta	-	St	P02254	194				
Strongylocentrotus purpuratus	H1-beta	Sp1be_P	P15869 85364	211 211				
Strongylocentrotus purpuratus	H1-beta	Sp1be_A Sp1be_N	AAA30052 NP_999723	211	Sp1be	M20314 NM_214558	1698 919 (633)	
Strongylocentrotus purpuratus	H1-delta	Sp1de	NP_999722	185	Sp1de	NM_214557	1138 (555)	
Strongylocentrotus purpuratus	H1-gamma	Sp1ga	NP_999720	217	Sp1ga	NM_214555	654 (651)	
Strongylocentrotus purpuratus	H1-early	Sp1early	NP_999714	205	Sp1early	NM_214549	715 (615)	
Sus scrofa	H1t	Ss1t	P06348 70666	211 211				
Tetraodon nigroviridis	-	Tn_CAF90	CAF90042	198				&

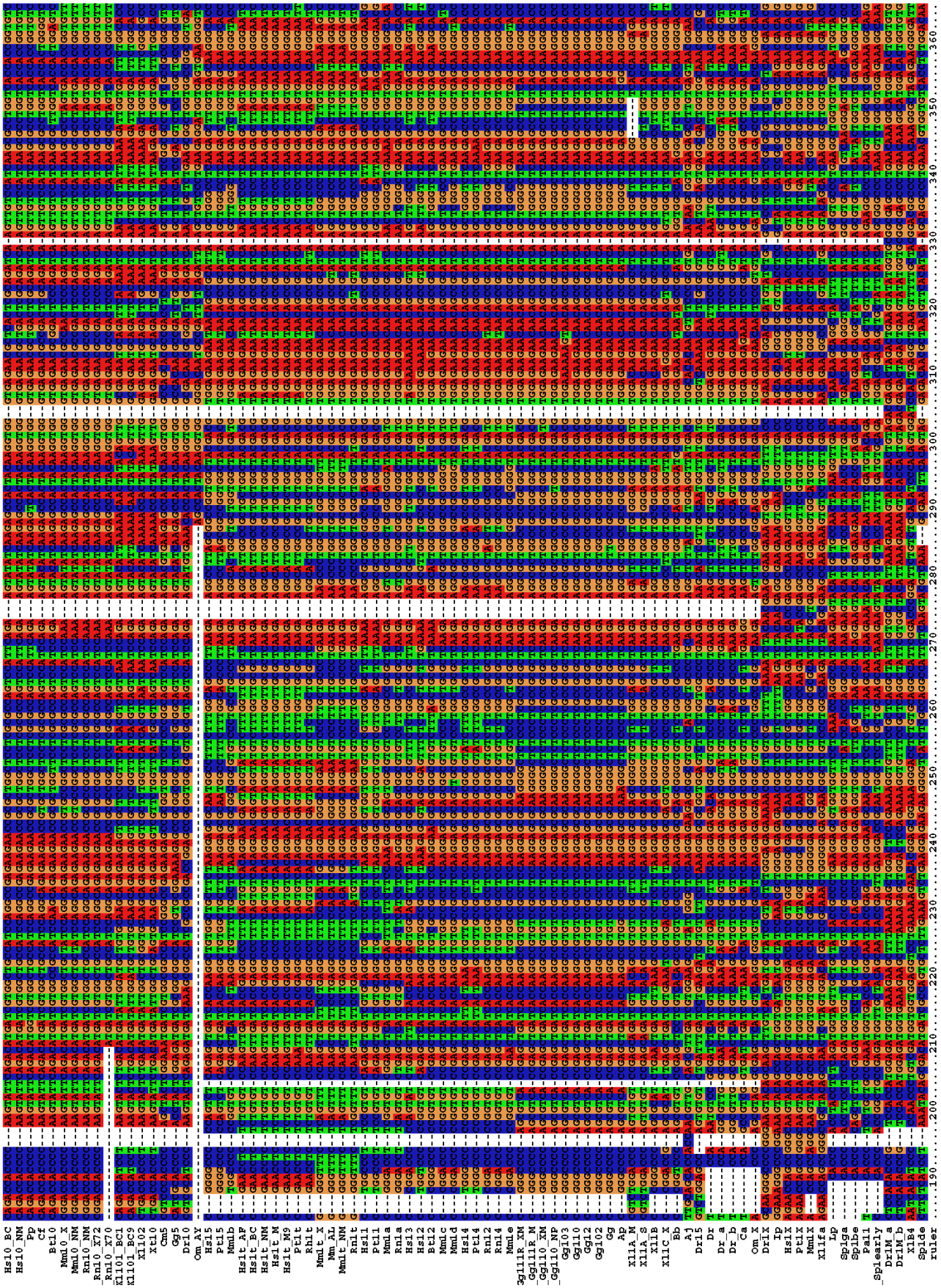
Tabela 5. Alfabetyczny spis sekwencji białkowych i nukleotydowych

Gatunek	Wariant histonu	Oznaczenie białka	Kod dostępu białka	Dłu-gość	Oznaczenie sekwencji nukleotydowej	Kod dostępu GenBank	Długość	
Tetraodon nigroviridis	-	Tn_CAF97	CAF97258	184				&
Tetraodon nigroviridis	-	Tn_CAG04	CAG04363	194				&
Xenopus tropicalis	$H1^0$	Xt10	AAH67985	196	Xt10	BC067985	2133 (588)	
Xenopus laevis	H1A	Xl1A_P	P06892 AAA49767 70669	210 210 209	Xl1A_M	M21287	8608 (630)	
Xenopus laevis	H1A	Xl1A_A	AAB29881	229	Xl1A_S	S69089	1022 (687)	
Xenopus laevis	H1B	Xl1B	P06893 AAA49764 2118972	220 220 219	Xl1B	M21286	660	
Xenopus laevis	H1C	Xl1C_C	CAA51433	221	Xl1C_X	XLH1CG	1793 (663)	
Xenopus laevis	H1C	Xl1C_P6	P15866	217				
Xenopus laevis	H1C	Xl1C_P7	P15867	221				
Xenopus laevis	H5B	Xl5B	P22845 85789	194 196				
Xenopus laevis	B4	HlB4	P15308	273	H1B4	X13855	1180 (819)	
Xenopus laevis	$H1^0$-1	Xl101_7	AAH72941	194	Xl101_BC1	BC072941	2229 (582)	
Xenopus laevis	$H1^0$-1	Xl101_5	AAH54149	176	Xl101_BC9	BC054149	1508 (528)	
Xenopus laevis	$H1^0$-2	Xl102	CAA96130	196	Xl102	Z71503	1959 (588)	
Xenopus laevis	H1X	Xl1fx	AAH41758	217	Xl1fa	BC041758	2269 (651)	

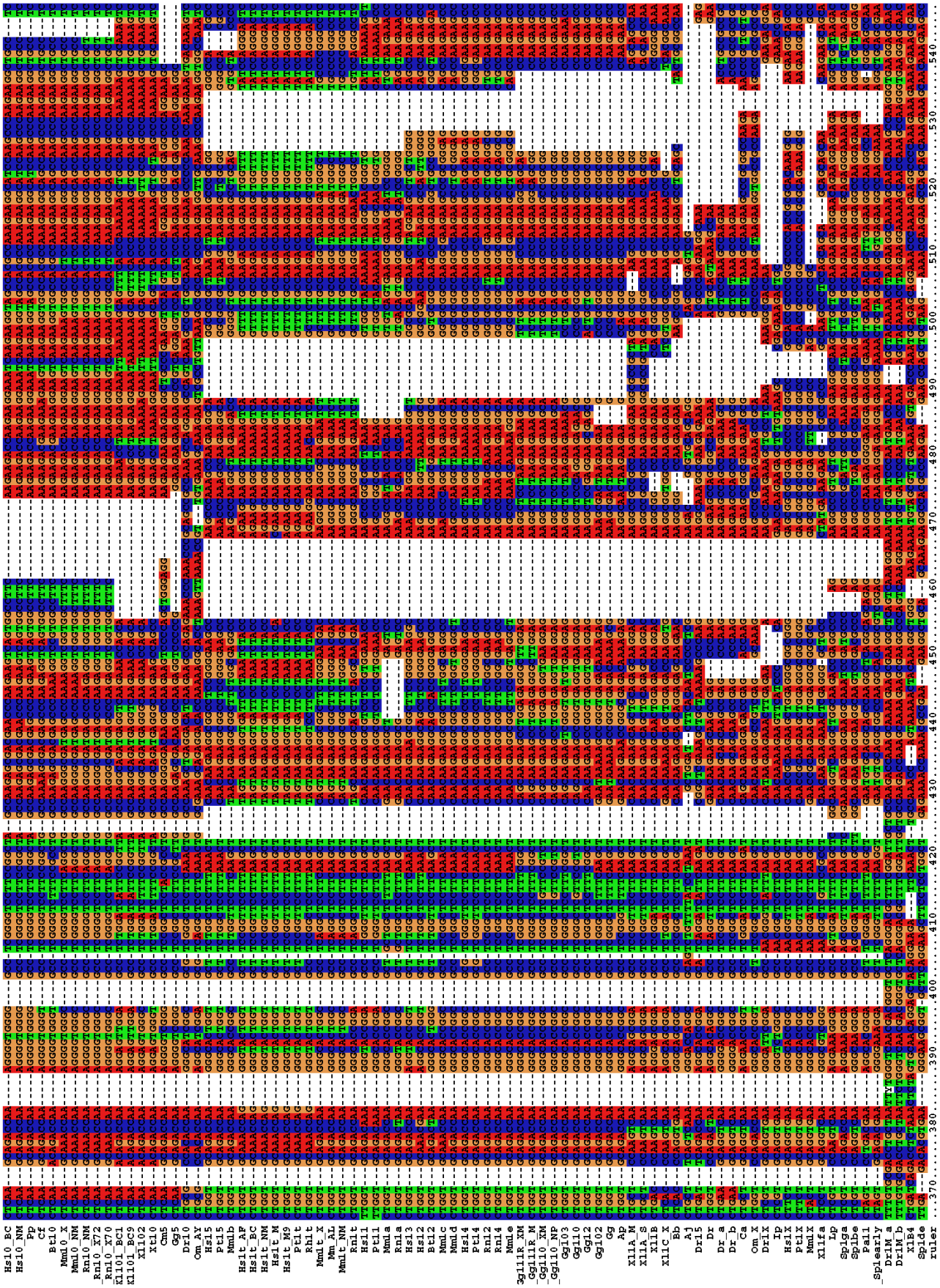
* sekwencja przewidziana metodami automatycznymi na podstawie DNA
& sekwencja nie opisana jako histon łącznikowy
( ) długość sekewncji kodującej



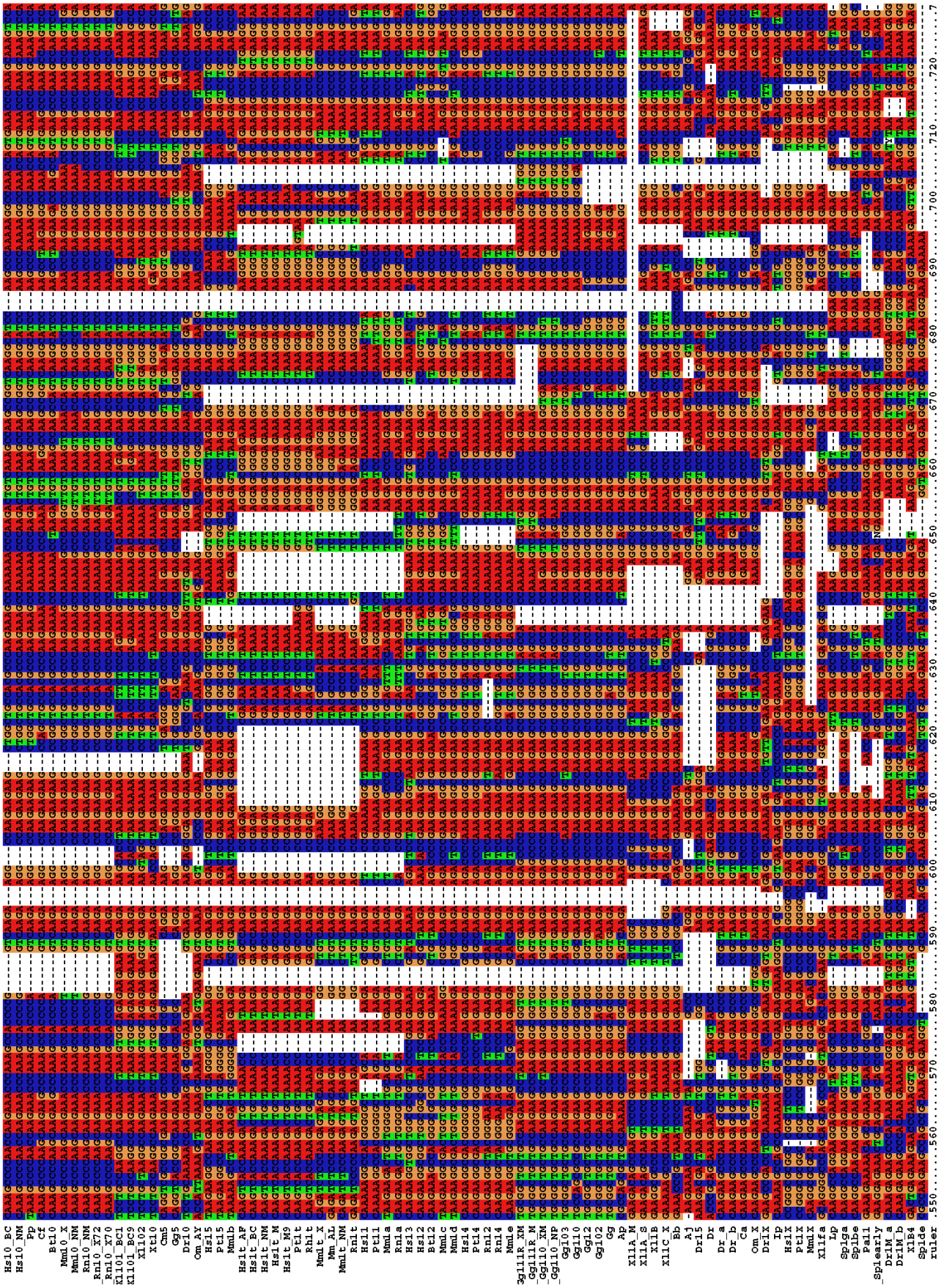
Rysunek 7. Zestawienie sekwencji nukleotydowych (MSA) histonów łącznikowych o długości 907 nukleotydów (pozycje 1-182). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6.



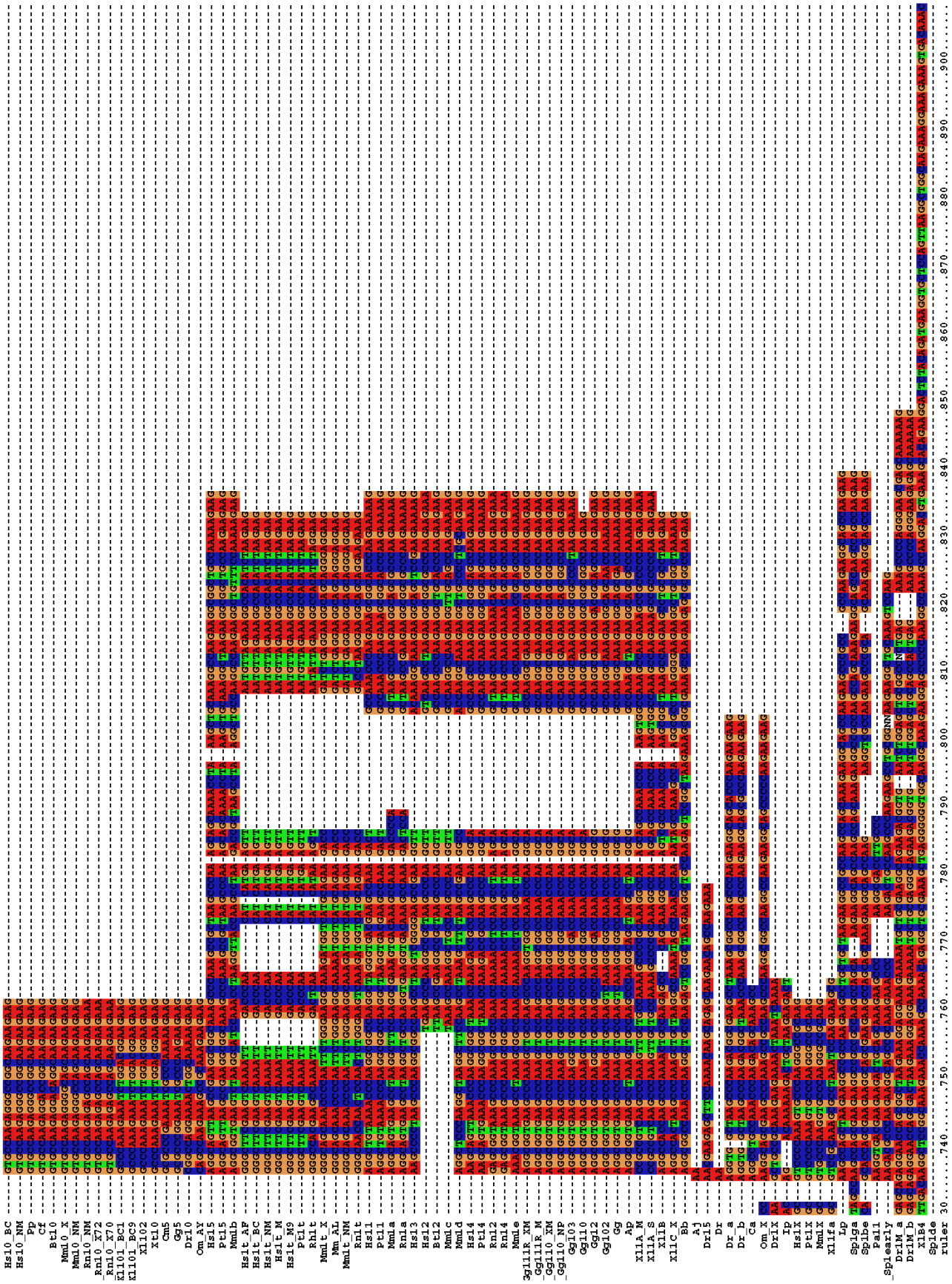
Rysunek 7. Zestawienie sekwencji nukleotydowych (MSA) histonów łącznikowych o długości 907 nukleotydów (pozycje 183-364). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.



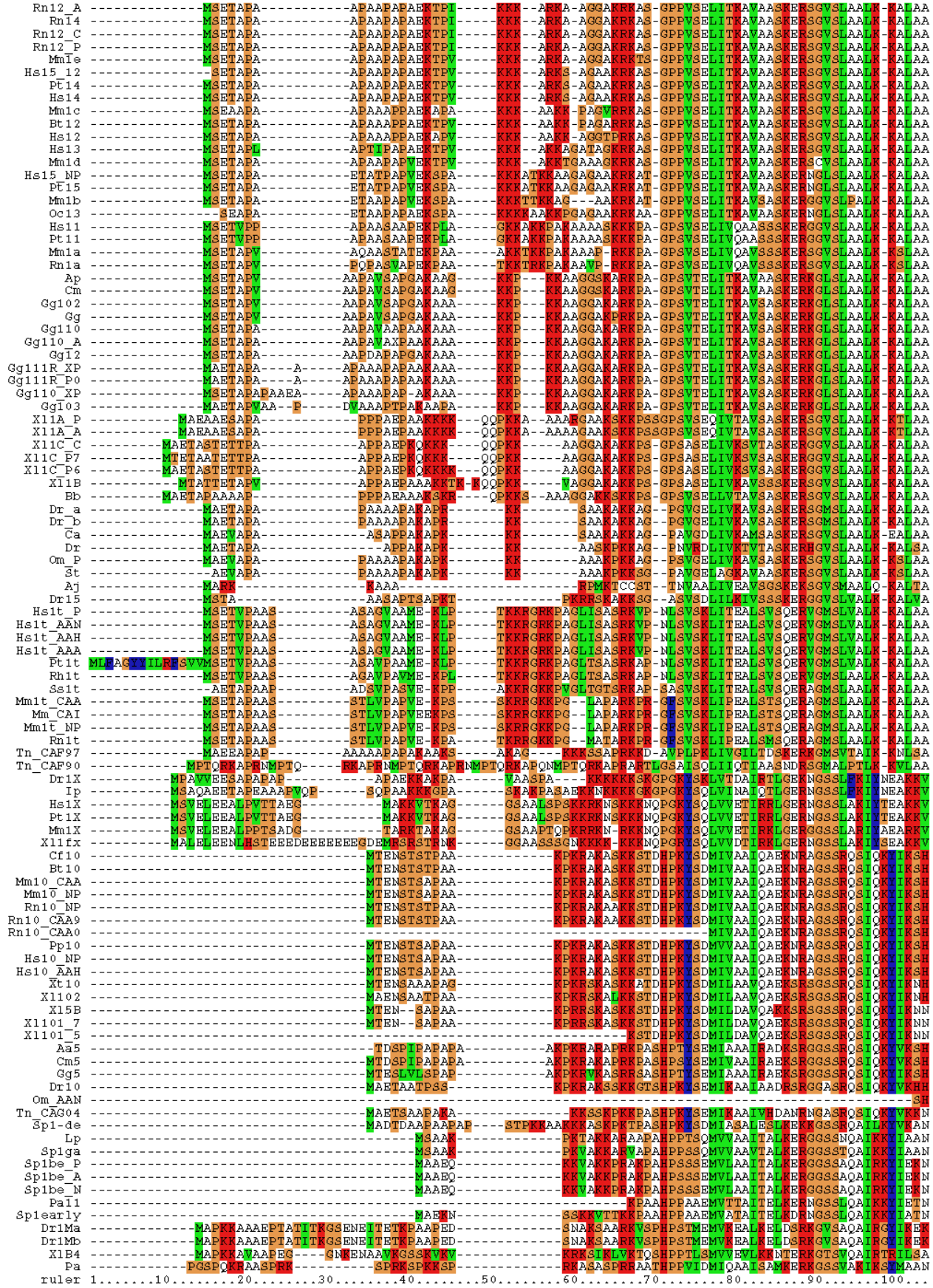
Rysunek 7. Zestawienie sekwencji nukleotydowych (MSA) histonów łącznikowych o długości 907 nukleotydów (pozycje 365-546). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.



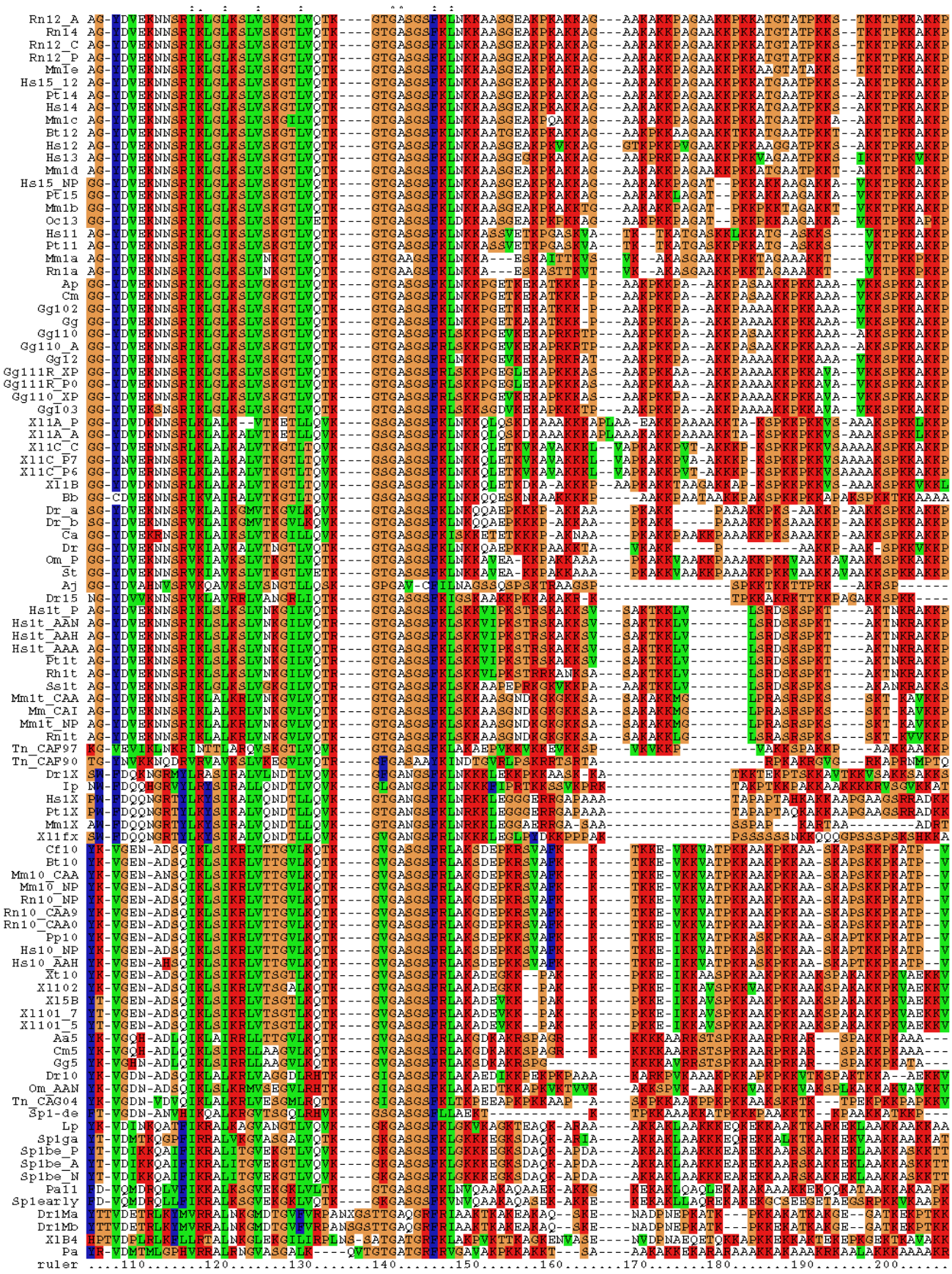
Rysunek 7. Zestawienie sekwencji nukleotydowych (MSA) histonów łącznikowych o długości 907 nukleotydów (pozycje 547-728). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.



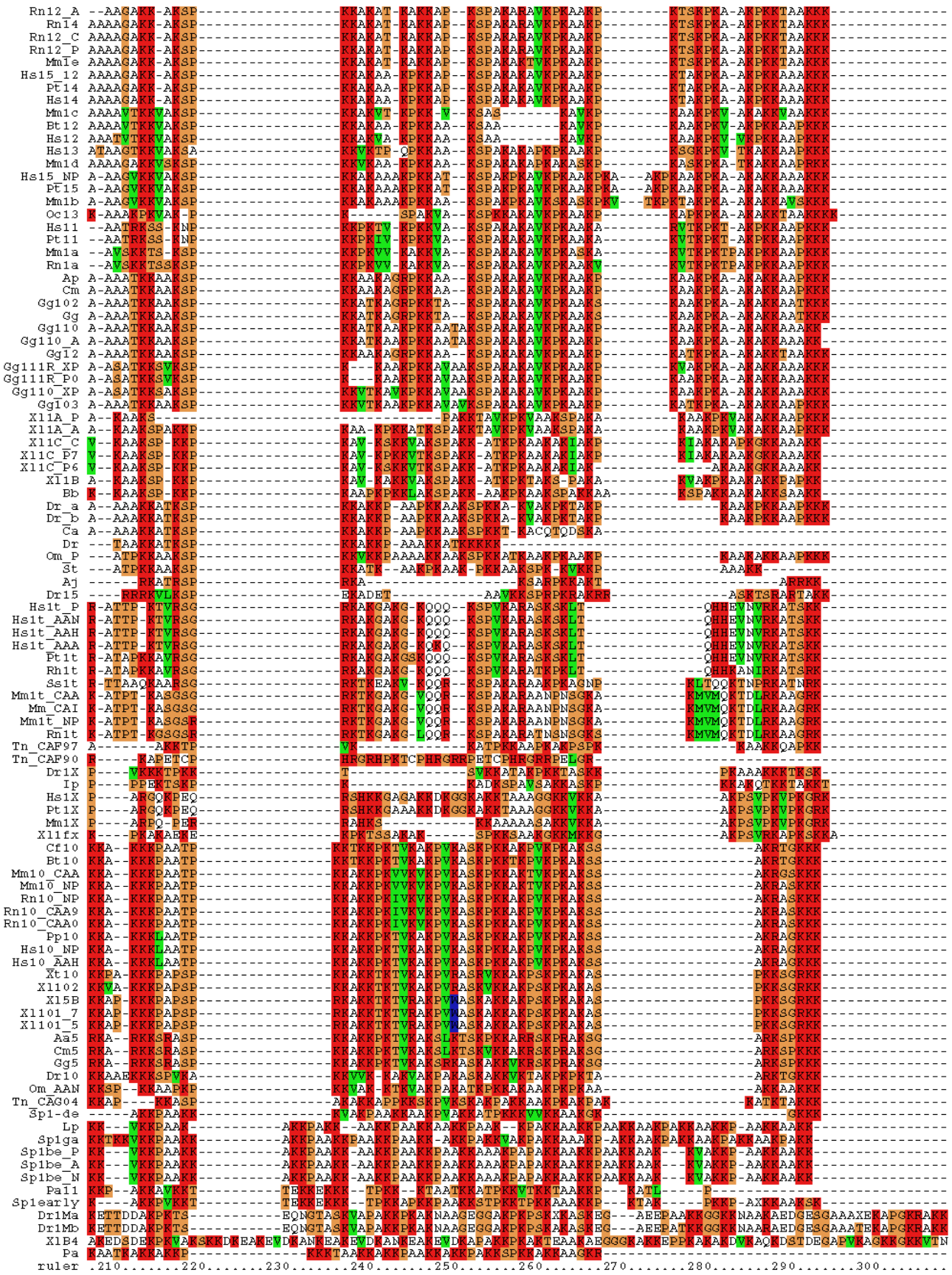
Rysunek 7. Zestawienie sekwencji nukleotydowych (MSA) histonów łącznikowych o długości 907 nukleotydów (pozycje 729-907). Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.



Rysunek 8. Zestawienie sekwencji białkowych (MSA) histonów łącznikowych o długości 308 aminokwasów (pozycje 1-102). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6.



Rysunek 8. Zestawienie sekwencji białkowych (MSA) histonów łącznikowych o długości 308 aminokwasów (pozycje 103-206). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.



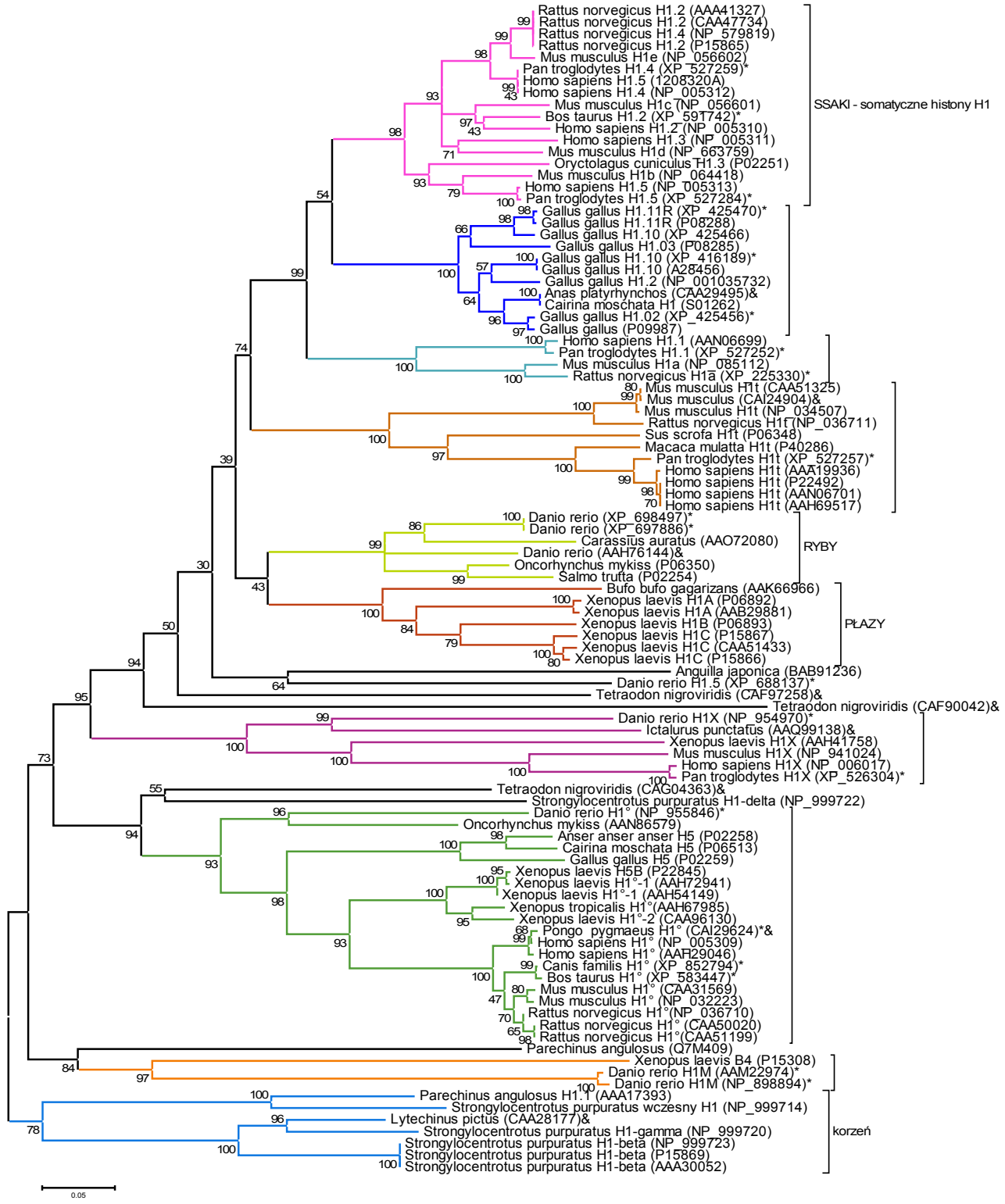
Rysunek 8. Zestawienie sekwencji białkowych (MSA) histonów łącznikowych o długości 308 aminokwasów (pozycje 207-308). Gwiazdkami (*) oznaczono miejsca konserwatywne, zaś myślniki (-) symbolizują przerwy. Oznaczenia sekwencji zgodne z tabelą 6. Ciąg dalszy.

4. WYNIKI

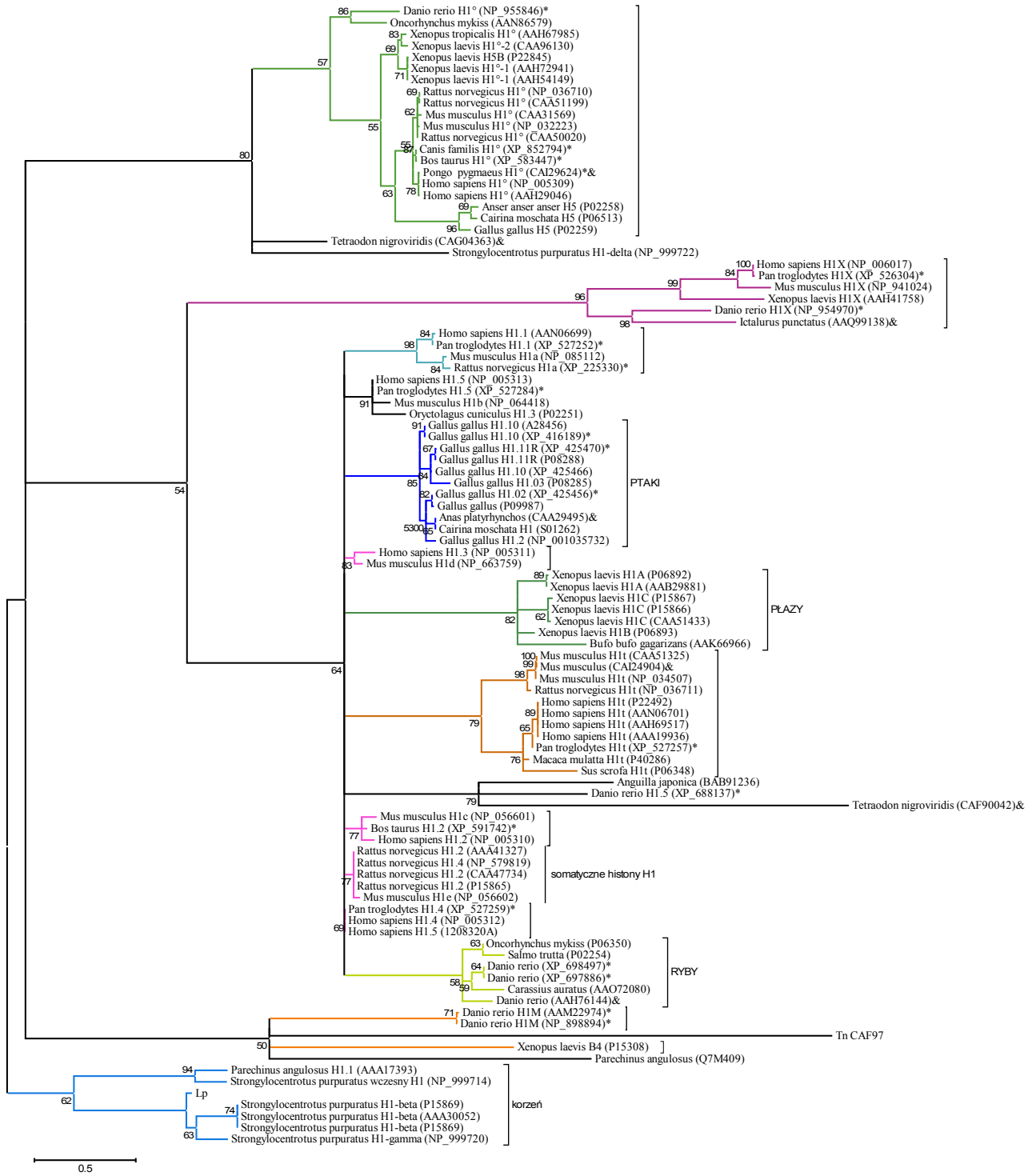
4.1 BIALKA HISTONOWE

Histony łącznikowe to podrodzina białkowa w obrębie histonów, która cechuje się najszybszym tempem ewolucji pozostając jednocześnie dość konserwatywną grupą białek. Histony H1 i H5 są małymi białkami o długości około 200 aminokwasów. Cechuje je wybitna zasadowość. Stosunek aminokwasów zasadowych do kwaśnych wynosi 7:1. Ponadto występuje tu prawie trzy razy więcej aminokwasów niepolarnych (Tabela 8). Histony łącznikowe zbudowane są z trzech domen z których najbardziej konserwatywna jest domena centralna, która wykazuje silne podobieństwo nawet u odległych ewolucyjnie organizmów. Widoczne to jest w zestawieniu białek histonowych, które w tym regionie posiada mało przerw oraz duże ogólne podobieństwo budujących je aminokwasów (Rysunek 7).

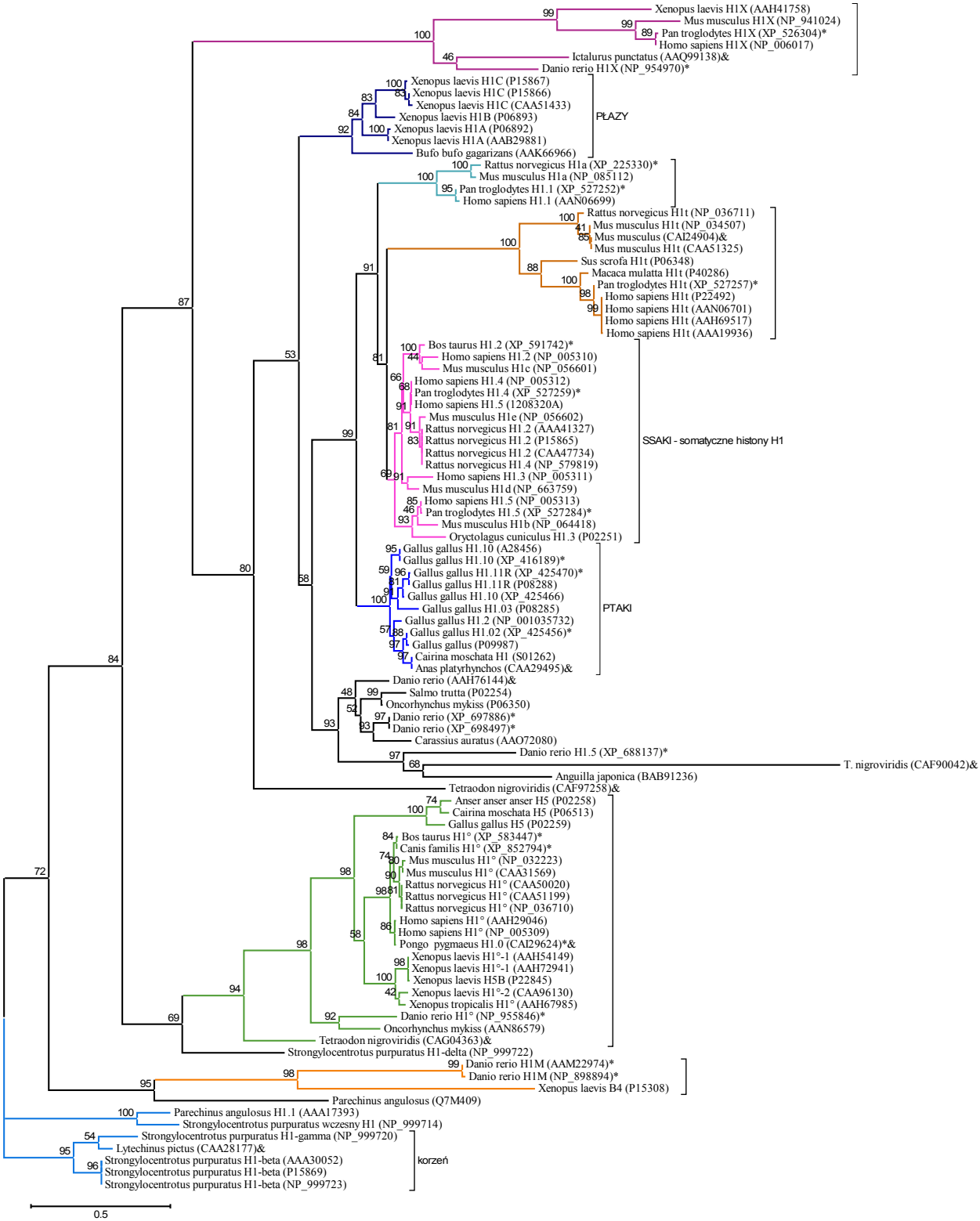
Ze względu na dużą liczbę krótkich sekwencji drzewo obrazujące stosunki filogenetyczne białek histonowych zostało wykonane w oparciu o niepoprawioną wartość p metodą NJ (Rysunek 9), ponieważ wartość ta ma małą wariancję i doświadczalnie wykazano, że w takich sytuacjach daje ona względnie dobre wyniki niejednokrotnie wyprzedzając inne bardziej skomplikowane metody (Nei i Kumar, 2000). Różnice odległości p pomiędzy poszczególnymi sekewencjami białkowymi wahały się w szerokim zakresie średnio wynosząc 0.44 ± 0.04. Dodatkowo skonstruowano drzewa w oparciu o metodę ML (Rysunek 10 i 11) oraz przeprowadzono mapowanie prawdopodobieństwa (Rysunek 12). Otrzymane drzewa nieznacznie się różnią, jednak każde z nich wykazuje charakterystyczną cechę polegającą na tym, że poszczególne białka grupują się w zależności od podtypu a nie ze względu na pochodzenie. Dopiero w obrębie poszczególnych grup można zauważyć typowe rozgałęzienia obrazujące relacje międzygatunkowe. Jedynym wyjątkiem od tej zasady są histony łącznikowe kury, które wyraźnie tworzą odrębną grupę. Na dzień dzisiejszy ciężko wynioskować czemu się tak dzieje, niezbędna będzie większa liczba danych pochodzących od innych ptaków.



Rysunek 9. Drzewo filogenetyczne białek histonowych H1 i H5 skonstruowane metodą najbliższego sąsiada NJ w oparciu o nieskorygowaną odległość p. Powyżej węzłów podano procentową wartość bootstrap (10 tys. powtórzeń). Obok nazwy gatunkowej i subtypu histonu w nawiasach podano kod dostępu GenBank. Gwiazdką oznaczono sekwencje przewidziane metodami automatycznymi. Znakiem & oznaczono sekwencje nie opisane w GenBank jako histony łącznikowe, ale mimo to umieszczone w Histone Data Base. Drzewo zbudowano za pomocą programu MEGA.



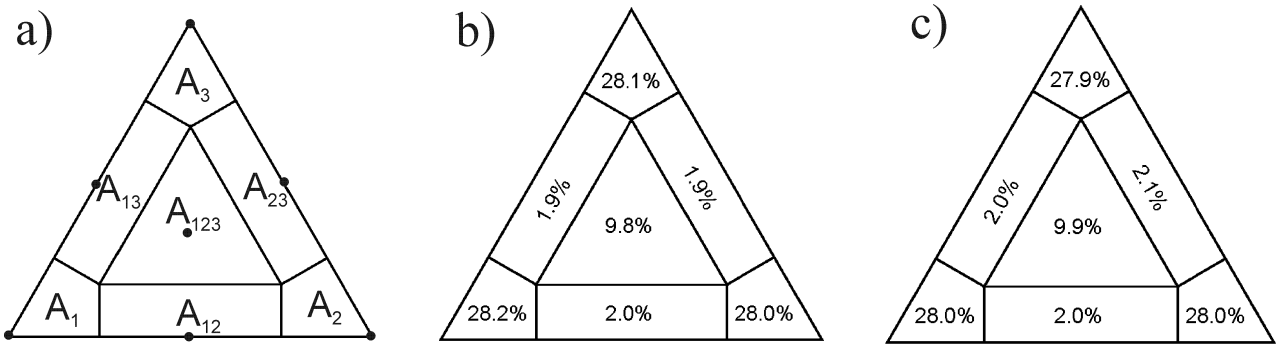
Rysunek 10. Drzewo filogenetyczne białek histonów łącznikowych skonstruowane metodą największego prawdopodobieństwa ML w oparciu o macierz Dayhoff z uwzględnieniem rozkładu gamma dla miejsc zmiennych (6 przedziałów) i miejsc niezmiennych. Powyżej węzłów podano procentową ilość kwartetów popierających dane rozgałęzienie (na podstawie 250 tys. losowo wybranych kwartetów). Drzewo zbudowano za pomocą programu TREE-PUZZLE. Oznaczenia jak na rysunku 9.



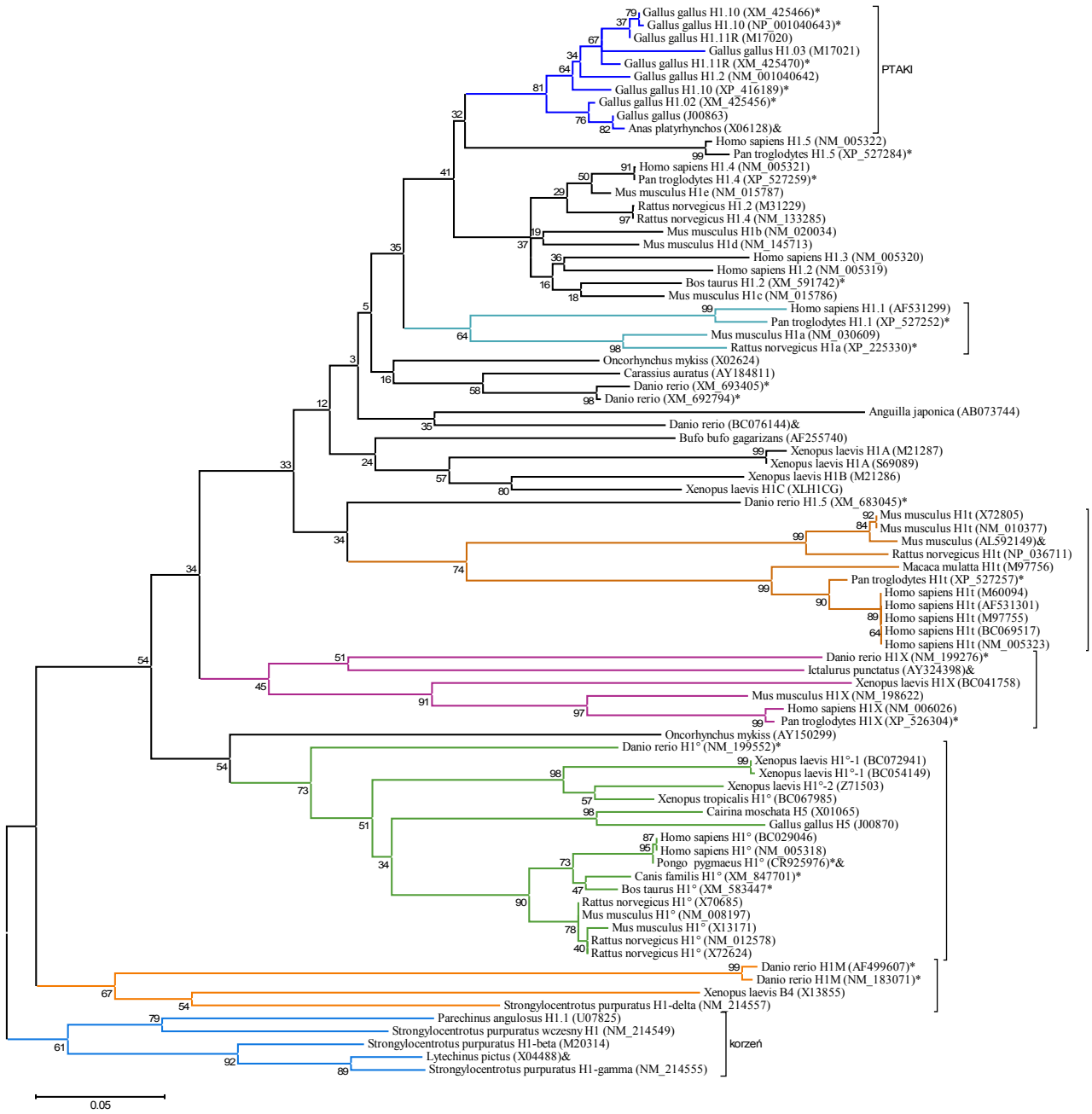
Rysunek 11. Drzewo filogenetyczne białek histonowych H1 i H5 skonstruowane metodą największego prawdopodobieństwa ML w oparciu o macierz Dayhoff z uwzględnieniem rozkładu gamma dla miejsc zmiennych (dyskretyzacja na 6 przedziałów) i miejsc niezmiennych. Powyżej węzłów podano procentową wartość bootstrap (1000 powtórzeń). Drzewo zbudowano za pomocą programu Treefinder. Oznaczenia jak na rysunku 9.

Tabela 7. Względny poziom użycia kodonów synomicznych (RSCU) genów histonów łącznikowych. Podano częstości występowania poszczególnych kodonów na podstawie, których wyliczono RSCU (wartości w nawiasach).

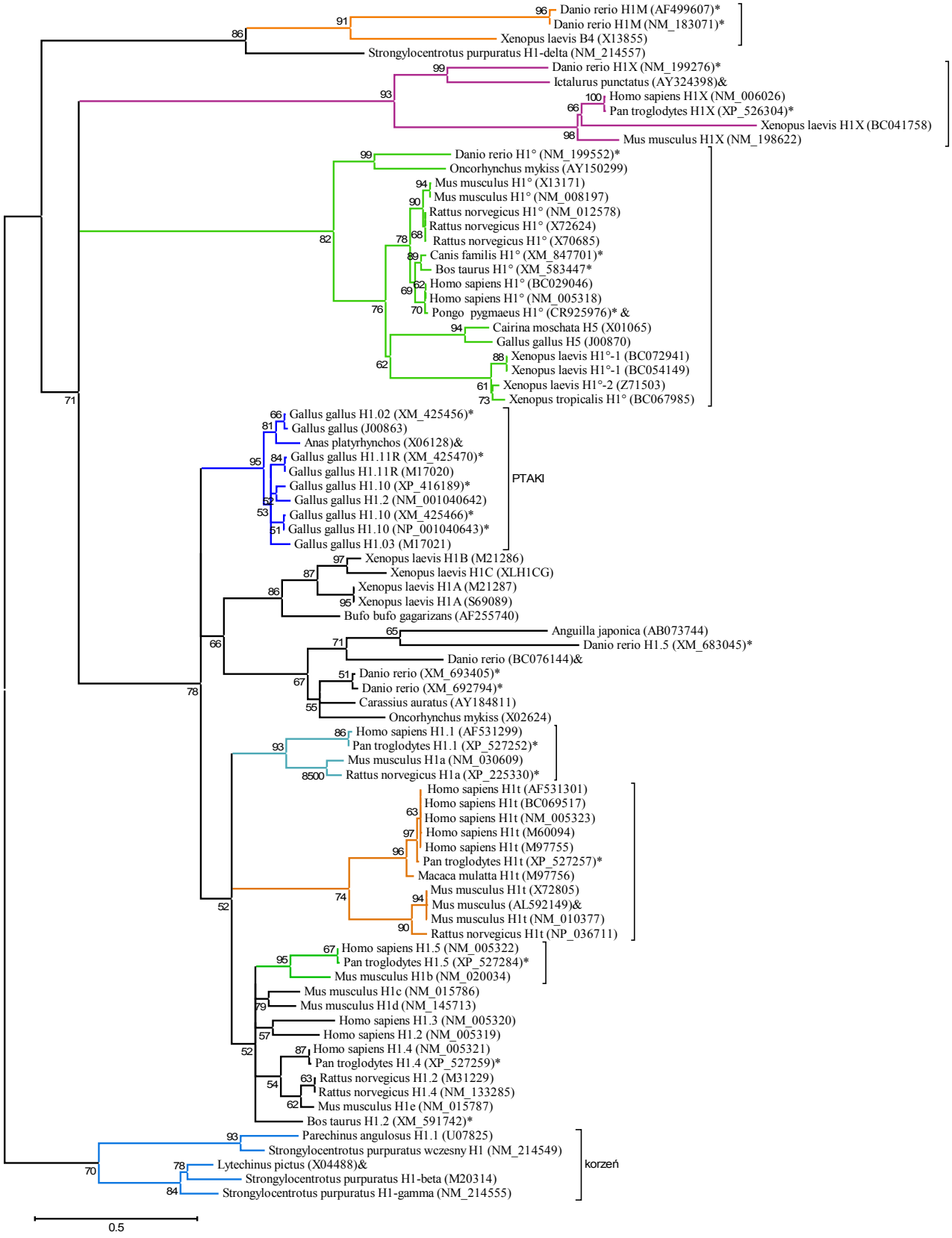
UUU(F)	0.4(0.36)	UCU(S)	1.3(0.44)	UAU(Y)	0.6(0.78)	UGU(C)	1.5(0.37)
UUC(F)	1.8(1.64)	UCC(S)	2.6(0.92)	UAC(Y)	0.9(1.22)	UGC(C)	6.4(1.63)
UUA(L)	0.3(0.36)	UCA(S)	0.9(0.33)	UAA(*)	4.3(2.20)	UGA(*)	1.2(0.60)
UUG(L)	0.0(0.02)	UCG(S)	0.5(0.18)	UAG(*)	0.4(0.21)	UGG(W)	1.7(1.00)
CUU(L)	0.6(0.80)	CCU(P)	3.3(1.43)	CAU(H)	1.6(0.74)	CGU(R)	1.6(0.61)
CUC(L)	2.4(3.19)	CCC(P)	3.2(1.40)	CAC(H)	2.7(1.26)	CGC(R)	6.7(2.54)
CUA(L)	0.9(1.19)	CCA(P)	1.5(0.64)	CAA(Q)	19.5(1.81)	CGA(R)	2.1(0.81)
CUG(L)	0.3(0.44)	CCG(P)	1.2(0.53)	CAG(Q)	2.0(0.19)	CGG(R)	2.5(0.93)
AUU(I)	0.3(0.40)	ACU(T)	1.0(0.72)	AAU(N)	0.7(0.37)	AGU(S)	2.5(0.88)
AUC(I)	2.2(2.48)	ACC(T)	3.7(2.49)	AAC(N)	3.2(1.63)	AGC(S)	9.2(3.24)
AUA(I)	0.1(0.12)	ACA(T)	0.6(0.39)	AAA(K)	6.5(1.49)	AGA(R)	1.5(0.58)
AUG(M)	0.7(1.00)	ACG(T)	0.6(0.40)	AAG(K)	2.2(0.51)	AGG(R)	1.4(0.52)
GUU(V)	0.8(0.60)	GCU(A)	3.7(0.99)	GAU(D)	1.4(0.56)	GGU(G)	4.0(0.63)
GUC(V)	3.4(2.40)	GCC(A)	7.3(1.95)	GAC(D)	3.5(1.44)	GGC(G)	15.8(2.48)
GUA(V)	0.7(0.48)	GCA(A)	2.2(0.58)	GAA(E)	17.9(1.59)	GGA(G)	2.8(0.44)
GUG(V)	0.7(0.52)	GCG(A)	1.8(0.48)	GAG(E)	4.6(0.41)	GGG(G)	2.9(0.45)



Rysunek 12. Mapowanie prawdopodobieństwa za pomocą trójkątów prawdopodobieństwa. a) lokalizacja w regionie A_1, A_2 i A_3 oznacza w pełni rozwiązaną strukturę w obrębie kwartetu; A_{12}, A_{13} i A_{23} dotyczy częściowo rozwiązanej struktury (dwa równie prawdopodobne rozwiązania) i centralny region A_{123} w którym znajdują się kwartety w których nie da się odtworzyć relacji filogenetycznych. b) trójkąt prawdopodobieństwa dla 99 białek histonów łącznikowych. c) trójkąt prawdopodobieństwa dla 83 sekwencji nukleotydowej histonów łącznikowych.



Rysunek 13. Drzewo filogenetyczne nukleotydowych sekwencji kodujących histony H1 i H5 skonstruowane metodą najbliższego sąsiada NJ na podstawie liczby synomicznych różnic między sekwencjami obliczonymi zmodyfikowaną metodą Nei-Gojobori (na podstawie odległości p oraz współczynnika R = 0.76). Powyżej węzłów podano procentową wartość bootstrap (10 tys. powtórzeń). Obok nazwy gatunkowej i subtypu histonu w nawiasach podano kod dostępu GenBank. Gwiazdką oznaczono sekwencje przewidziane metodami automatycznymi. Znakiem & oznaczono sekwencje nie opisane w GenBank jako histony łącznikowe, ale mimo to umieszczone w Histone Data Base. Drzewo zbudowano za pomocą programu MEGA.



Rysunek 14. Drzewo filogenetyczne ML nukleotydowych sekwencji kodujących histony H1 i H5 skonstruowane metodą kwartetów według modelu HKY z rozkładem gamma dla miejsc zmiennych (6 przedziałów) i miejsc niezmiennych. Powyżej węzłów podano procentową ilość kwartetów popierających dane rozgałęzienie (na podstawie 250 tys. losowo wybranych kwartetów). Drzewo zbudowano za pomocą programu TREE-PUZZLE. Oznaczeń jak na rysunku 9.

Rysunek 15. Drzewo filogenetyczne ML sekwencji nukleotydowych kodujących histony łącznikowe w oparciu o model HKY z rozkładem gamma dla miejsc zmiennych (6 przedziałów) i miejsc niezmiennych. Powyżej węzłów podano procentową wartość bootstrap (1000 powtórzeń). Drzewo zbudowano za pomocą programu Treefinder. Oznaczenia jak na rysunku 9.

Tabela 8. Procentowy udział aminokwasów budujących białka histonów łącznikowych (średnia z 99 znalizowanych sekwencji). Kolorem czerwonym zaznaczono aminokwasy kwaśne, niebieskim aminokwasy zasadowe, zielonym aminokwasy polarne.

Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu
20.37	0.05	1.20	3.17	0.67	5.70	0.41	1.77	26.46	4.34
Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
0.91	1.81	8.16	1.81	3.55	8.32	5.14	5.30	0.04	0.82

4.2 SEKWENCJE NUKLEOTYDOWE

Analizę sekwencji nukleotydowych ograniczono do regionów kodujących białka. Wynika to z fragmentaryczności danych, w wielu przypadkach brak jest pełnych odcinków promotorowych. Zbadano stosunek poszczególnych zasad, który przedstawia się: T – 0.11755, C – 0.28604, G – 0.28838, A – 0.30802. Dodatkowo obliczono względny poziom użycia kodonów synomicznych RSCU (Tabela 7). Współczynnik tranzycji/transwersji R wynosi 0.76 ± 0.03 (odpowiednio 0.77 ± 0.04, 0.93 ± 0.07, 0.66 ± 0.03 dla pierwszej, drugiej i trzeciej pozycji kodonów). Zbadano różnicę synomiczne (p_S) i niesynomiczne (p_N) między poszczególnymi sekwencjami. W cele zbadania jaki rodzaj selekcji kieruje ewolucją histonów łącznikowych przeprowadzono test statystyczny Z oparty na kodonach. Jako hipotezę zerową H_0 przyjęto $p_S = p_N$ (selekcja neutralna), a hipotezami alternatywnymi H_1 i H_2 były $p_S > p_N$ (selekcja oczyszczająca) i $p_S < p_N$ (selekcja pozytywna). Hipoteza zerowa została odrzucona z prawdopodobieństwem błedu $P < 0.05$. Prawdziwą okazała się być jedynie hipoteza H_1.

Podobnie jak w przypadku sekwencji białkowych skonstruowano trzy drzewa filogenetyczne według podobieństwa sekwencjach nukleotydowych. Pierwsze oparte na względnie prostych metodach dystansu między sekwencjami (Rysunek 13) i dwa pozostałe zbudowane według metod ML (Rysunek 14 i 15). Dodatkowo przeprowadzono mapowanie prawdopodobieństwa za pomocą programu TREE-PUZZLE (Rysunek 12b).

5. DYSKUSJA

Białka histonowe są rodziną silnie konserwatywnych białek wsród których wyróżnia się pięć podrodzin. Jedną z nich są histony łącznikowe H1 i H5. Jest to heterogenna grupa złożona z kilku białek, których liczba zależy od organizmu i może wynosić do 9 w zależności od gatunku. Ponadto występują one w wielu kopiach. Fakt ten potwierdzają badania organizmów pochodzących z różnych królestw według których histony łącznikowe są obecne zawsze przynajmniej w dwóch formach u wszystkich zanalizowanych pod tym kątem organizmów (ponad 100 gatunków grzybów, roślin i zwierząt; Kasinsky i wsp., 2001).

Celem niniejszych badań nie jest ustalenie kiedy poszczególne formy histonów łącznikowych powstały, ponieważ według badań Ponte i współpracowników somatyczne histony łącznikowe i histon H1t rozdzieliły się 406 ± 80 milionów lat temu czyli wcześniej niż wykształciły się ssaki, a nawet kręgowce (Ponte i wsp., 1998). Podstawowym zadaniem pracy jest zbadanie jak kształtował się proces ewolucji histonów łącznikowych i jaki rodzaj selekcji był główną siłą kształtującą relacje w tej rodzinie. Poznanie tych procesów pod kątem molekularnym pozwoli rozstrzygnąć jak będzie się kształtować ten proces w dalszej przyszłości.

Ewolucję wielogenowych rodzin starano się wytłumaczyć za pomocą kilku modeli ewolucyjnych. Pierwszym takim modelem był model ewolucji różnicującej (divergent evolution) według którego poszczególne geny powstały z macierzystego genu w wyniku procesu duplikacji i stopniowego nagromadzania się różnic między powstałymi genami. Model ten doskonale tłumaczył proces ewolucji białek rodziny globulinowej jednak w miarę napływu nowych danych okazał się nieadekwatny do innych rodzin białkowych w których zaobserwowano, że często mamy doczynienia z wieloma tandemowo powtórzonymi genami, które są niemal identyczne u danego organizmu i różne u organizmów pokrewnych. Taka sytuacja kształtuje się na przykład w obrębie rodziny rybosomalnych RNA (rRNA). W związku z tym zaproponowano nowy model nazwany ewolucją jednorodną (concerted evolution). Zakłada on, że wszystkie geny rodziny ewoluują w sposób jednorodny jako jedna

niepodzielna całość czyli poszczególne geny rodziny przestają być niezależne. Wszelkie zachodzące mutacje rozprzestrzeniają się na sąsiednich przedstawicieli rodziny za pomocą wielokrotnego crossing over lub w wyniku konwersji genów. Aż do początku lat 90-tych większość badaczy przyjmowała, że właśnie tak kształtują się procesy ewolucji rodzin genowych. Wszelkie sprzeczności, które pojawiały się wraz z nagromadzeniem coraz większej ilości danych molekularnych próbowno wyjaśnić odpowiednio modyfikując model. Jednak zastosowanie ewolucji jednorodnej nie było możliwe w przypadku olbrzymiej różnorodności z jaką się zetknięto w przypadku białek układu odporności a mianowicie przeciwciał i białek MHC. W związku z powyższym zaproponowano inny, obowiązujący do dziś, model ewolucji narodzin i śmierci (birth-and-death evolution). Zakłada on powstanie nowych genów w wyniku duplikacji w efekcie których nowe geny mogą zostać włączone na stałe lub też mogą być usunięte lub przekształcone w pseudogeny (Nei i Hughes, 1992; Nei i Rooney, 2005). To właśnie obecność pseudogenów była przyczyną głównych kontrowersji i źródłem krytyki modelu ewolucji jednorodnej. Istnieje możliwość rozróżnienia, który model ewolucji lepiej pasuje do analizowanych danych. Jednym podejściem może być zbadanie jaki rodzaj selekcji dominuje. Jeśli będziemy mieli doczynienia z selekcją oczyszczającą najprawdopodobniejszym modelem będzie model birth-and-death. Z drugiej strony należy przyjrzeć się relacjom jakie panują między poszczególnymi członkami rodziny.

Podstawowym założeniem selekcji oczyszczającej jest to, że liczba różnic synomicznych (p_S) przewyższa liczbę różnic niesynomicznych (p_N). Tak też kształtuje się sytuacja w przypadku histonów łącznikowych. Przeprowadzony test Z oparty na kodonach jednoznacznie wykluczył pozostałe dwie możliwości. Niemniej jednak w przypadku niewielkich różnic między analizowanymi genami test ten może okazać się zbyt liberalny. W istocie porównanie między kilkoma genami pewnych subtypów histonów łącznikowych występujących u jednego organizmu (np. H1t u *Homo sapiens*) nie pozwala rozstrzygnąć tego problemu. W takich przypadkach o ewolucji pewnych białek nie da się wnioskować w ten sposób. Problem ten można rozwiązać inaczej. Należy podejść do sprawy od drugiej strony i sprawdzić jakie są

konsekwencje ewolucji zgodnie z założeniami poszczególnych modeli, a następnie porównać je z otrzymanymi wynikami. Właśnie takie podejście zostało wykorzystane w niniejszej pracy. Jeśli przyjmiemy, że rodzina genowa ewoluuje zgodnie z modelem jednorodnym powinniśmy oczekiwać, że poszczególni członkowie rodziny będą bardzo podobni w obrębie gatunku, a różni między gatunkami. Efektem tego będzie skupienie się genów danego gatunku w jedną grupę na drzewie filogenetycznym. Inaczej będzie wyglądać sytuacja w przypadku ewolucji birth-and-death. Tutaj geny będą się skupiać w grupy pod względem przynależności do danego subtypu a nie gatunku.

Analizę histonów łącznikowych przeprowadzono na poziomie nukleotydów i białka. Otrzymane wyniki są bardzo podobne i bez względu na zastosowaną metodę konstrukcji drzew wykazują wyraźne podobieństwa. Na początku skonstruowano drzewa w oparciu o metody dystansu (Rysunek 9 i 13) stosując algorytm NJ. W pierwszym przypadku podstawą drzewa była nieskorygowana odległość p, ponieważ białka histonowe są krótkie, a liczba analizowanych sekwencji była dość duża. W takim przypadku uzasadnione wydaje się zastosowanie nawet tak uproszczonego modelu ewolucji. Dla zobrazowania procesów ewolucji sekwencji nukleotydowych zastosowano zmodyfikowaną metodę Nei-Gojobori, która rozróżnia substytucje synomiczne od niesynomicznych oraz pozwala uwzględnić różną częstość zachodzenia tranzycji i transwersji (współczynnik R). Metodę tą wybrano, ponieważ wartość R = 0.76 różniła się od wartości oczekiwanej w przypadku jednakowego tempa tranzycji i transwersji (R = 0.5). Metody te mają dodatkową zaletę jaką jest mała liczba parametrów co pozwala na szybkie otrzymanie rezultatów (czas konstruowania tych drzew nie przekraczał 10 minut na komputerze Celeron 2.6GHz). Następnym etapem badań było skonstruowanie drzew w oparciu o zasady największego prawdopodobieństwa. W przypadku białek zastosowano macierz Dayhoff (Tabela 4). Ponadto za pomocą rozkładu gamma uwzględniono możliwość nierównomiernego tempa substytucji w zależności od znaczenia danego miejsca dla funkcjonowania genu. Wprowadzono także dodatkowy paramatr dla miejsc konserwatywnych ewolucyjnie (Rysunek 10 i 11). Analizę sekwencji

nukleotydowych oparto o model HKY (Tabela 3), który bierze pod uwagę współczynnik R oraz różną proporcję poszczególnych zasad (Rysunek 14 i 15). Otrzymane topologie różniły się nieznacznie nawet w przypadku zastosowania metod zaliczanych do tej samej grupy (ML). Wynikać to może z kilku czynników. Otóż w przypadku metod ML zastosowano dwa odmienne programy znacznie różniące się podejściem do problemu. Program TREE-PUZZLE opiera się na dość szczególnej metodzie kwartetów. Pozwala ona wnioskować o stosunkach filogenetycznych na podstawie kwartetów (czwórek) sekwencji i tak otrzymane wyniki ekstrapoluje się tworząc drzewo końcowe. Dodatkową zaletą programu jest to, że przed rozpoczęciem głównej analizy można sprawdzić za pomocą mapowania prawdopodobieństwa czy dane są odpowiednio dobrane. W przypadku sekwencji histonów łącznikowych wyniki te były wystarczająco dobre (ponad 80% kwartetów faworyzuje ściśle określoną topologię). Jednak blisko w 10% przypadków metoda kwartetów okazała się być niejednoznaczna. Procent ten był identyczny w przypadku sekwencji białkowych i nukleotydowych co jest dość dziwne, ponieważ z założenia im dłuższe sekwencje tym lepiej powinna sprawować się matoda kwartetów. Jedynym wytłumaczeniem może być to, że w obrębie analizowanej grupy sekwencji część z nich była nieodpowiednia wprowadzając zakłócenia. Najwyższy procent nierozstrzygniętych kwartetów posiadały krótkie sekwencje białkowe (na przykład histon H1 pochodzący z *Anguilla japonica*) oraz długie histony podtypu H1M/B4. Mimo to nie zostały one usunięte. Prawdopodobnie to jest przyczyną powstania licznych węzłów wyższego rzędu. Niemniej jednak drzewa otrzymane za pomocą tego programu nadal wyglądają dobrze. Dodatkowo zbudowano drzewa ML za pomocą programu Treefinder, ponieważ symulacje potwierdziły jego wyższość nad programem TREE-PUZZLE. Program ten skutecznością dorównuje programom z pakietu PHYLIP będąc wiele razy szybszym. Dlatego za najbliższe rzeczywistości uznaję właśnie drzewa skonstruowane za pomocą programu Treefinder. Dodatkowo w tym przypadku test bootstrap daje wysokie poparcie dla większości otrzymanych gałęzi.

Bez względu na metodę konstrukcji drzew oraz model ewolucji, który przyjęto

wyniki okazały się być wyjątkowo spójne. Histony łącznikowe grupują się w zależności od podtypu danego białka/genu. Dopiero w obrębie grupy można zaobserwować typowe relacje międzygatunkowe jakich można by oczekiwać pod kątem pokrewieństwa. Rozdział na grupy jest silnie poparty testem bootstrap i jest obecny w każdym z drzew. Jedynym wyjątkiem od tej zasady są histony łącznikowe ptaków i płazów, które wyraźnie tworzą oddzielne grupy skupiające wszystkie podtypy z wyjątkiem histonu H1° i H5. Przyczyny tego zjawiska nie są znane i jego wyjaśnienie będzie wymagać większej liczby sekwencji pochodzących od ptaków i płazów, które w tej chwili nie są dostępne. Jako pierwsza wyodrębnia się grupa histonów H1M/B4, następnie powstają histony niezależne od procesu replikacji H1° i H5 oraz histony H1X. Kolejną grupą są somatyczne histony łącznikowe wraz z histonami H1t. W obrębie tej grupy jako pierwsze oddzielają się histony H1t. Spośród histonów somatycznych najlepiej wyróżniającą się grupą jest podtyp H1.1. Pomimo tych podobieństw trafiają się istotne różnice między poszczególnymi drzewami. Analizą objęto także szereg sekwencji otrzymanych metodami automatycznymi oraz nie opisanych jako histony łącznikowe, ale mimo to umieszczonymi w Histone Data Base. Większość z nich ustawicznie lokuje się w tym samym miejscu względem innych sekwencji co pozwala jednoznacznie zaliczyć je do określonego podtypu (na przykład histon H1° pochodzący od *Bos taurus* i *Canis familis* w przypadku sekwencji przewidzianych metodami automatycznymi oraz białko *Tetraodon nigroviridis* oznaczone jako CAG04363). Czasem sekwencje takie okazały się jedynie zbliżonymi do histonów łącznikowych o czym może świadczyć nienormalnie wydłużona gałąź tak jak to ma miejsce w przypadku białka *Tetraodon nigroviridis* (CAF90042), które prawdopodobnie histonem łącznikowym nie jest.

Otrzymane wyniki jednoznacznie wskazują że rodzina histonów łącznikowych ewoluuje zgodnie z modelem birth-and-death z silną selekcją oczyszczającą. Podobny wniosek otrzymali także inni autorzy (Eirin-Lopez i wsp., 2004; Eirin-Lopez i wsp., 2005), jednak ich badania opierały się głównie na porównaniu proporcji substytucji synomicznych (p_S) i niesynomicznych (p_N), a zastosowane metody konstrukcji drzew były ograniczone jedynie do metod dystansu. W niniejszej

pracy potwierdzono te wyniki stosując wiele innych metod, które na dzień dzisiejszy uznaje się za najodpowiedniejsze i najbardziej miarodajne. Należy podkreślić, że również ewolucja histonów rdzeniowych kształtowana jest zgodnie z modelem birth-and-death (Piontkivska i wsp., 2002; Rooney i wsp., 2002).

WAŻNIEJSZE ADRESY INTERNETOWE

GenBank	http://www.ncbi.nlm.nih.gov/genbank
Histone Sequence Database	http://research.nhgri.nih.gov/histones/
Treeview	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
RasMol	http://www.umass.edu/microbio/rasmol/index2.htm
Treefinder	http://www.treefinder.de/
ClustalX	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
ClustalW	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/
TREE-PUZZLE	http://www.tree-puzzle.de/
PHYLIP	http://evolution.genetics.washington.edu/phylip.htm
MEGA 3.1	http://www.megasoftware.net/
Prottest	**http://darwin.uvigo.es/software/prottest.html**
Findmodel	http://hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html

LITERATURA

1. Alami, R., Fan, Y., Pack, S., Sonbuchner, T.M., Besse, A., Lin, Q., Graelly, J.M., Skoultchi, A.I., Bouhassira, E.E. (2003) Mammalian linker-histone sybtypes differentially affect gene expression *in vivo*. Proc. Natl. Acad. Sci. 100(10), 5920-5925.
2. Rooney, A.P., Piontkivska, H., Nei, M. (2002) Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. Mol. Biol. Evol. 19(1), 68-75.
3. Nei, M., Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39, 121-152.
4. Piontkivska, H., Rooney, A.P., Nei, M. (2002) Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. Mol Biol Evol 19, 689-697.
5. Eirin-Lopez, J.M., Gonzalez-Tizon, A.M., Martinez, A., Mendez, J. (2004) Birth-and-Death Evolution with Strong Purifying Selection in the Histone H1 Multigene Family and the Origin of orphon H1 Genes. Mol Biol Evol 21(10), 1992-2003.
6. Eirin-Lopez, J.M., Ruiz, M.F., Gonzalez-Tizon, A.M., Martinez, A., Ausio, J., Sanchez, L., Mendez, J. (2005) Common evolutionary origin and birth-and-death process in the replication-independent histone H1 isoforms from vertebrate and invertebrate genomes. J Mol Evol 61, 398-407.
7. Nei, M., Hughes, A.L. (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In *11th Histocompatibility Workshop and Conference*, ed. Tsuji, K., Aizawa, M., Sasazuki, T. pp. 27–38. Oxford, UK: Oxford Univ. Press.
8. Holder, M. T. (2001) Using a Complex Model of Sequence Evolution to Evaluate and Improve Phylogenetic Methods. PhD thesis. The University of Texas at Austin, USA.

9. Hedges, S.B., Kumar, S., Tamura, K. (1992) Human orgins and analysis of mitochondrial DNA sequences. Science 255, 737-739.
10. Baxevalis, A.D. I Ouellette, B.F.F. (2004) Bioinformatyka. Podręcznik do analizy genów i białek. Wydawnictwo Naukowe PWN, Warszawa.
11. Bednar, J., Horowitz, R.A., Grigoriev, S.A., Carruthers, L.M., Hansen, J.C.,Koster, A.J., Woodcock, C. L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc. Natl. Acad. Sci. USA 95, 14173-14178.
12. Chakravarthy, S., Park, Y.J., Chodaparambil, J., Edayathumangalam, R.S., Luger, K. (2005) Structure and dynamic properties of nucleosome core particles. FEBS Lett. 579, 895-898.
13. Clarke, H.J., Oblin, C., Bustin, M. (1992) Developmental regulation of chromatin composition during mouse embryogenesis: somatic histone H1 is first detectable at the 4-cell stage. Development 115, 791-799.
14. Cotton, J.A. (2003) Vertebrate phylogenomics and gene family evolution. PhD thesis. University of Glasgow. Scotland.
15. Dominsky, Z., Marzluff, W.F. (1999) Formation of the 3`end of histone mRNA. Gene 239, 1-14.
16. Dong, Y., Liu, D., Skoultchi, A.I. (1995) An upstream control region required for inducible transcription of the mouse H1º histone gene during terminal differentiation. Mol. Cell. Biol. 15(4), 1889-1990.
17. Fan, Y., Nikitina, T., Morin-Kensicki, E.M., Zhao, J., Magnuson, T.R., Woodcock, Ch.L., Skoultchi, A.I. (2003) H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. Mol. Cell. Biol. 23(13), 4559-4572.
18. Fan, Y., Sirotkin, A., Russell, R.G., Ayala, J., Skoultchi, A.I. (2001) Individual somatic H1 subtypes are dispensable for mouse development even in mice lacking the H1º replacement subtype. Mol. Cell. Biol. 21(23), 7933-7943.
19. Fan, Y., Skoultchi, A.I. (2003) Genetic analisis of H1 linker histone subtypes

and their functions in mice. Methods Enzymol., 377, 85-107.

20. Fantz, D.A., Hatfield, W.R., Horvath, G., Kistler, M.K., Kistler W.S. (2001) Mice with targeted disruption of the H1t gene are fertile and undergo normal changes in structural chromosomal proteins during spermatogenesis. Biol. Reprod. 64, 425-431.
21. Felsenstein, J. (2004) Inferring Phylogenies. Sinauer Associates.
22. Franke, K., Drabent, B., Doenecke, D. (1998) Expression of murine H1 histone genes during postnatal development. Biochim. Biophys. Acta 1398, 232-242.
23. Gadagkar, S.R., Kumar, S. (2005) Maximum Likelihood Outperforms Maximum Parsimony Even When Evolutionary Rates Are Heterotachous. Mol. Biol. Evol. 22(11), 2139-2141
24. Gajiwala, K.S., Burley, S.K. (2000) Winged helix proteins. Curr. Opin. Struct. Biol. 10, 110-116.
25. Grimes, S.R., Wilkerson, D.C., Noss, K.R., Wolfe, S.A. (2003) Transcriptional control of the testis-specyfic histone H1t gene. Gene 304, 13-21.
26. Hendzel, M.J., Level, M.A., Crawford, E., Th`ng, J.P.H. (2004) The Cterminal domain is the primary determinant of histone H1 binding to chromatin in vivo. J. Biol. Chem. 279(19), 20028-34.
27. Henikoff, S. i Henikoff J.G. (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Res. 19, 6565-6572.
28. Henikoff, S. i Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915-10919.
29. Jenuwein, T., Allis, D.C. (2001) Translating the histone code. Science 293(5532), 1074-1080.
30. Jobb, G., von Haeseler, A., Stimmer, K. (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evolutionary Biology 4, 18-26.
31. Kasinsky, H.E., Lewis, J.D., Dacks, J.B., Ausio, J. (2001) Origin of H1 linker histones. The FASEB Journal 15, 34-42.
32. Khochbin, S. (2001) Histone H1 diversity: binding regulatory signals to linker

histone function. Gene 271, 1-12.

33. Khochbin, S. i Wolffe, A.P. (1994) Developmentally regulated exspressin of linker-histone variants in vertebrates. Eur. J. Biochem. 225, 501-510.
34. Khorasanizadeh, S. (2004) **T**he nucleosome from genomic organization to genomic regulation. Cell 116, 259–272.
35. Kłyszejko – Stefanowicz, L. (2002). Cytobiochemia. Biochemia niektórych struktur komórkowych. Wydawnictwo Naukowe PWN, Warszawa.
36. Koutzamani, E., Loborg, H., Sarg, B., Lindner, H.H., Rundquist, I. (2002) Linker histone subtype composition and affinity for hromatin *in situ* in nucleated mature eryhrocytes. J. Biol. Chem. 277(47), 44688-44694.
37. Kowalski, A., Pałyga, J., Górnicka-Michalska, E. (2004) Identification of histone H1.z components in a Muscovy duck (*Cairina moschata* L.) population. Comp. Biochem. Physiol. B 137, 151-157.
38. Kozłowski, Ł. (2004) Zróżnicowanie histonu H1 u kręgowców. Akademia Świętokrzyska. Praca licencjacka.
39. Kumar, S. (1996) A Stepwise Algorithm for Finding Minimum Evolution Trees. Mol. Biol. Evol. 13(4), 584-593.
40. Kumar, S., Tamura, K., Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. 5, 150-163
41. Lemmon, A.R., Milinkovitch, M.C. (2002) The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. Proc. Natl. Acad. Sci. USA 99, 10516-10521.
42. Lennox, R.W. i Cohen, L.H. (1983) The histone H1 complements of dividing and nondividing cells of the mouse. J. Biol. Chem. 258(1), 262-268.
43. Lin, Q., Sirotkin, A., Skoultchi, A.I. (2000) Normal spermatogenesis in mice lacking the testis-specyfic linker histone H1t. Mol. Cell. Biol. 20(6), 2122-2128.
44. Luger, K., Hansen, J.C. (2005) Nucleosome and chromatin fiber dynamics. Curr. Opin. Struct. Biol. 15, 188-196.

45. Nei, M., Kumar, S. (2000) Molecular evolution and phylogenomics. Oxford University Press.
46. Pałyga, J. (1990) Variability of histone H1 in rabbit populations. Int. J. Biochem. 22, 1351-1361.
47. Pałyga, J., Górnicka-Michalska, E., Kowalski, A., Książkiewicz, J. (2000) Natural allelic variation of duck eryhrocyte histone H1b. Int. J. Biochem. Cell Biol. 32, 665-675.
48. Piontkivska, H. (2004) Efficiencies of maximum likelihood methods of phylogenetic interferences when different substitutions model are used. Mol. Phylogent. Evol. 31, 865-873.
49. Ramakrishnan, V. (1997) Histone structure and the organization of the nucleosome. Annu. Rev. Biophys. Biomol. Struct. 26, 83-112.
50. Saitou, N. i Nei, M. (1987) The Neighbor-joining Method: A New Method forReconstructing Phylogenetic Trees. Mol. Biol. Evol. 4(4), 406–425.
51. Sirotkin, A.M., Edelmann, W., Cheng, G., Klein- Szanto, A., Kucherlapati, Skoultchi, A.I. (1995) Mice develop normally without the H1º linker histone. Proc. Natl. Acad. Sci. 92, 6434-6438.
52. Smith, A.D., Lui T.W.H., Tillier E.R.M. (2004) Empirical models for substitution in ribosomal RNA. Mol. Biol. Evol. 21(3), 419–427.
53. Sonnhammer, E.L.L., Hollich, V. (2005) *Scoredist:* A simple and robust protein sequence distance estimator. BMC Bioinformatics 6, 108-11.
54. Ponte, I., Vidal-Taboada, J.M., Suau, P. (1998) Evolution of the vertebrate H1 histone class: evidence for the functional differentation of the subtypes. Mol. Biol. Evol. 15(6), 702-708.
55. Strimmer, K.S. (1997) Maximum likelihood methods on molecular phylogenetics. PhD thesis. University of Munich. Germany.
56. Strimmer, K., von Haeseler, A. (1996) Quartet-puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13, 964–969.
57. Sullivan, J. (2005) Maximum likelihood methods for phylogeny estimation.

Method. Enzymol. 395, 757-779.

58. Sullivan, S., Sink, D.W., Trout, K.L., Makalowska, I., Baxevanis, A.D., Landsman, D. (2002) The Histone Database. Nucleic Acids Res. 30(1), 341-342.

59. Takami, Y., Nakayama, T. (1997b) A single copy of linker H1 genes is enogh for proliferation of the DT40 chicken B cell line, and linker H1 variants participate in regulation of gene expression. Genes Cells 2, 711-723.

60. Takami, Y., Nishi, R., Nakayama, T. (2000) Histone H1 variants play individual roles on transcription regulation in the DT40 chicken B cell line. Biochem. Biophys. Res. Commun. 268, 501-508.

61. Tanaka, M., Hennebold, J.D., Macfarlane, J., Adashi, E.Y. (2001) A mammalian oocyte-specific linker histone gene H1oo: homology with the genes for the oocyte-specific cleavage stage histone (cs-H1) of sea urchin and the B4/H1M histone of the frog. Development 128, 655-664.

62. Tanaka, M., Hennebold, J.D., Macfarlene, J., Adashi, E.Y. (2001) A mammalian oocyte-specyfic linker histone H1oo: Homology with the genes for the oocyte-specyfic cleavage stage histone (cs-H1) of sea urchin and the B4/H1M histone of the frog. Development 128, 655-664.

63. Thompson, J.D., Gibson, T.J., Higgins, D.G. (2003) Multiple sequence alignment using ClustalW and ClustalX. Curr. Prot. Bioinf. 2.3.1-2.3.22

64. Travers, A. (1999) The location of the linker histone on the nucleosome. Trends Biochem. Sci. 24, 4-7.

65. Turner, B.M. (2002) Cellular memory and histone code. Cell 111, 285–291.

66. Vila, R., Ponte, I., Collado, M., Arrondo, J.L.R., Jiménez, M.A., Rico, M., Suau, P. (2001) DNA-induced α-Helical Structure in the NH2-terminal Domain of Histone H1. J. Biol. Chem. 276(49), 46429-46435.

67. Vila, R., Ponte, I., Jiménez, M.A., Rico, M., Suau, P. (2000) A helix-turn motif in the C-terminal domain of histone H1. Protein Sci. 9, 627-636.

68. Vila, R., Ponte, I., Jiménez, M.A., Rico, M., Suau, P. (2002) An inducible helix–Gly–Gly–helix motif in the N-terminal domain of histone H1e. A CD

and NMR study. Protein Sci. 11, 214-220.

69. Wang, Z.-F., Sirotkin, A.M., Buchold, G.M., Skoultchi, A.I., Marzluff, W.F., (1997) The mouse histone H1 genes: gene organization snd differential regulation. J. Mol. Biol. 271, 124-138.
70. Whelan, S., Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18, 691-699.
71. Wierzbicki, A.T. (2002) Zagadka histonu H1. Post. Biochem. 43 (3), 167-174
72. Wierzbicki, A.T., Jerzmanowski, A. (2004) Suppression of histone H1 genes in Arabidopsis results in heritable developmental defects and stochastic changes in DNA methylation. Genetics 104.031997v1.
73. Wilkerson, D.C., Wolfe, S.A., Grimes, S.R. (2002) H1t/GC-box and H1t/TE1element are essential for promoter activity of the testis-specyfic histone H1t gene. Biol. Reprod. 67, 1157-1164.
74. Yap, V.B., Speed, T. (2005) Rooting a phylogenetic tree with nonreversible substitution models. BMC Evolutionary Biology 5, 2.
75. Bernstein, H.J. (2000) Recent changes to RasMol, recombining the variants. Trends Biochem. Sci. 25, 453-455.
76. Abascal, F., Zardoya, R., Posada, D. (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics: 21(9), 2104-2105.
77. Posada, D., Crandal, K.A. (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9), 817-818.

Kielce, dnia 20.06.2006

Łukasz Kozłowski
/Imię i nazwisko/

67668
/Numer albumu/

OŚWIADCZENIE

Świadomy odpowiedzialności karnej oświadczam, że przedkładana praca magisterska

pt.: Analiza filogenetyczna histonów łącznikowych kręgowców.

została napisana przeze mnie samodzielnie oraz nie narusza praw autorskich zgodnie z Ustawą z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (tekst jednolity: Dz.U. z 2000 roku Nr 80, poz. 904 z późn. zm.).

Wyrażam zgodę
na udostępnienie mojej pracy licencjackiej
dla celów naukowych i dydaktycznych.

..……..……...
/Podpis autora pracy/